Analyzing Dynamic Adversarial Training Data in the Limit

Anonymous ACL submission

Abstract

To create models that are robust across a wide range of test inputs, training datasets should include diverse examples that span numerous phenomena. Dynamic adversarial data collection (DADC), where annotators craft examples that challenge continually improving models, holds promise as an approach for generating such diverse training sets. Prior work has shown that running DADC over 1–3 rounds can help models fix some error types, but it does not necessarily lead to better generalization beyond adversarial test data. We argue that running DADC over *many* rounds maximizes its training-time benefits, as the different rounds can together cover many of the task-relevant phenomena. We present the first study of longer-term DADC, where we collect 20 rounds of NLI examples for a small set of premise paragraphs, with both adversarial and non-adversarial approaches. Models trained on DADC examples make 26% fewer errors on our expert-curated test set compared to models trained on non-adversarial data. Our analysis shows that DADC yields examples that are more difficult, more lexically and syntactically diverse, and contain fewer annotation artifacts compared to non-adversarial examples.

1 Introduction

Traditional crowdsourcing methods often yield datasets that lack diversity, contain spurious correlations, and are easy for existing models (Gururangan et al., 2018; Poliak et al., 2018; Geva et al., 2019; Ko et al., 2020; Potts et al., 2021). Training on such examples can lead to models that reach deceptively high accuracy on in-distribution test data, yet fail on challenge sets (Naik et al., 2018; Glockner et al., 2018; Gardner et al., 2020), input perturbations (Wallace et al., 2019; Kaushik et al., 2020), and distribution shifts (Talmor and Berant, 2019; Hendrycks et al., 2020).

Dynamic adversarial data collection (DADC) holds promise as an approach to mitigate these training set problems. In DADC, humans are tasked with creating examples that fool state-of-the-art models but are answerable by humans. Crucially, DADC is dynamic in that data collection is repeated over many rounds with a stream of ever-improving models-in-the-loop. As models improve, annotators are incentivized to craft new types of examples that challenge the latest models. In the limit, this process would ideally cover most task-relevant phenomena, leading to more robust models.

Whether DADC actually leads to diverse, high-coverage training data, however, has remained unclear. It could cause annotators to write unnatural examples or to focus on a narrow subset of unusual examples that models find difficult to learn, thus decreasing data diversity (Bowman and Dahl, 2021).

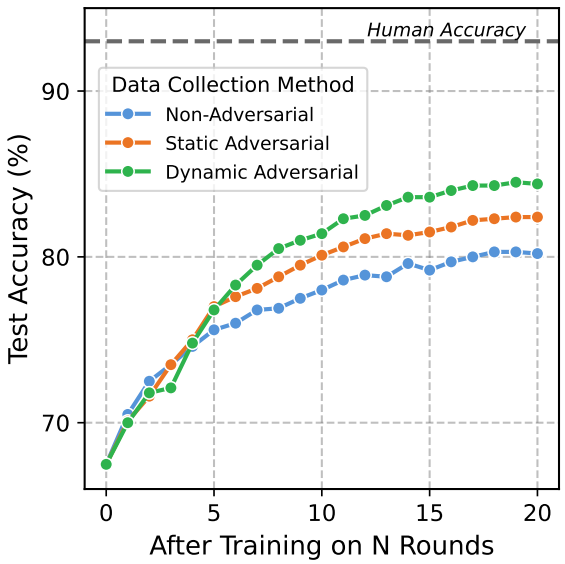

Figure 1: *Model accuracy on our expert-curated test set* when training on data collected from three different methods. Standard non-adversarial data collection is more effective than adversarial data collection in the short-term. However, in the long term, adversarial data collection statistically significantly outperforms standard data, especially when the data is collected using a dynamic model that is updated after each round.

Some prior work has shown that a few rounds of DADC can indeed improve robustness to adversarial inputs (Dinan et al., 2019; Nie et al., 2020a), however, there are mixed results on improving accuracy on other distributions (Kaushik et al., 2021). To date, no study has analyzed how DADC evolves over *many* rounds. Thus, the long-term benefits or drawbacks of adopting it as a core dataset creation paradigm remain poorly understood.

In this work, we conduct the first study of DADC's effects in the long term, where we conduct many rounds and rapidly update models. We focus on the task of natural language inference (NLI), which serves as a crucial benchmark for research on language understanding (Bowman et al., 2015; Williams et al., 2018a). To make our study feasible, we conduct intensive data collection on a small set of context passages that span different genres and exhibit numerous natural language phenomena. By using a small set of contexts, we create a scenario in which models can improve quickly from round to round, thus approximating the dynamics of running DADC at a larger scale. We compare three approaches for collecting training data—no model, static model-in-the-loop, and dynamic model-in-the-loop—in a controlled setting for 20 rounds.

To evaluate the different methods, we collect expert-curated non-adversarial test examples for each context that span numerous NLI phenomena which humans can handle correctly. On this test set, DADC outperforms the alternative approaches after many rounds of data collection (e.g., Figure 1). Standard non-adversarial data collection causes model accuracy to climb quickly for a short period of time, but accuracy quickly plateaus after more examples are collected. On the other hand, DADC examples yield larger improvements for later rounds. To understand these results, we show that DADC examples are overall more diverse in lexical and syntactic patterns, contain fewer artifacts, and become more difficult over each round. Overall, our results show that building large adversarial training sets may be more useful than standard model-agnostic collection in the long term.

2 Background

Collecting Data with Crowdsourcing. Most large-scale supervised datasets are collected using crowd workers (Bowman et al., 2015; Rajpurkar et al., 2016; Kočiský et al., 2018). Compared to experts, crowd workers often produce lower quality data as they are not necessarily well-trained for one's task and can be apathetic to the goals of the research (Snow et al., 2008; Gadiraju et al., 2017). These data quality issues are exacerbated for language tasks because crowd workers also need to *write* inputs, e.g., writing hypothesis sentences for natural language inference tasks. These manually-written inputs often follow a very narrow distribution: they lack diversity over lexical items, syntactic patterns, domains, example difficulties, reasoning types, and more (Yang et al., 2018; Gururangan et al., 2018; Geva et al., 2019; Min et al., 2019; Kiela et al., 2021).

Dynamic Adversarial Data Collection. In DADC, workers are tasked with writing examples that are answerable by humans but fool existing models (Wallace et al., 2019; Nie et al., 2020a; Kiela et al., 2021). Concretely, workers are presented with a user interface where they can observe model predictions and interactively build data that exposes model failures. Multiple *rounds* may also be conducted, where the model is updated on the adversarial data collected thus far and redeployed; the goal of this is to encourage workers to write increasingly more difficult examples. Adversarial data collection has been widely adopted in recent work, especially for building *evaluation* datasets (Dua et al., 2019; Nie et al., 2020a; Dinan et al., 2019; Bartolo et al., 2020; Potts et al., 2021; Liu et al., 2021; Kaushik et al., 2021; Xu et al., 2020, 2021). Our focus is instead on *training*, where past work has shown that after a few rounds of adversarial data, a model noticeably improves on its errors, yet many problems still remain (Nie et al., 2020a; Bartolo et al., 2020; Kaushik et al., 2021; Zellers et al., 2019). Moreover, it remains unclear whether collecting adversarial or non-adversarial data leads to generally more robust models in the long term (Kaushik et al., 2021).

3 Dynamic Data Collection in the Limit

The paradigm of DADC raises a natural but unanswered question: what would happen if we kept going? If we ran DADC for many years, how robust would the resulting models be? Would models improve more quickly than if we had collected training data without a model-in-the-loop?

Answering these forward-looking questions is key to understanding whether researchers and practitioners should continue to collect data in an adversarial fashion. Of course, we cannot practically

Premise	Model	Rd	Hypotheses	Label	Error
Sound	No	20	Old telephones have sheepskin over a cup or cylinder.	Entail	-
	Static	20	Parts of animal anatomy can function as the origins of sound.	Entail	✗
	Dynamic	20	The transmission due to the vibration can be attenuated with distances.	Entail	✓
Yellow	No	20	Ruiz's experiment was on three men.	Contradict	-
	Static	20	It turned out that basset hounds were immune to yellow fever.	Contradict	✓
	Dynamic	20	The American Public Health Association meeting, held in October 1900, was about developing vaccines against yellow fever.	Contradict	✗
Faraday	No	20	michael faraday's mother was named margaret	Entail	-
	Static	20	The home of the Faradays, in London, was very crowded.	Entail	✗
	Dynamic	20	Michael had at least nine uncles and/or aunts.	Entail	✓

Table 1: Examples from the training sets that are generated by crowd workers, with *No*, *Static*, or *Dynamic* models in the loop. The error column shows whether the worker successfully fooled the model in the loop when submitting the example in the user interface. See Table 5 for the full premise paragraphs.

run many years of data collection at once due to cost and time constraints. Our key idea is to instead answer these questions for a more manageable test bed that still retains many of the key challenges associated with language understanding tasks. In particular, we scale down the natural language inference (NLI) task to a small number of paragraph-length premises. In this setting, many rounds of smaller-scale data collection can tell us whether DADC or non-adversarial data collection leads to more robust model accuracy on test hypotheses for these same contexts. If DADC is indeed superior, this suggests that DADC can collect data that more effectively covers the challenging phenomena required for NLI, and therefore scaling it up to (many) more contexts could yield models that are similarly robust for more general NLI.

3.1 Task and Context Paragraphs

We choose to focus on NLI, a canonical and well-studied natural language understanding task (Dagan et al., 2005; Bos and Markert, 2005; Giampiccolo et al., 2007; MacCartney and Manning, 2009). NLI training datasets are notorious for being rife with artifacts and biases (Poliak et al., 2018; Gururangan et al., 2018; Tsuchiya, 2018; McCoy et al., 2019b), which makes NLI a suitable test bed for studying questions surrounding training dataset quality. Using NLI also enables us to write a rich and diverse test set with a small number of contexts because each premise admits many possible hypotheses. We focus on *binary* NLI—definitely entailing or not entailing—to minimize labeling disagreements stemming from the distinction between neutral and contradiction in three-way NLI (Pavlick and Kwiatkowski, 2019; Nie et al., 2020b).

We use ten diverse paragraphs from Project Gutenberg[1] as the premises—each one is chosen to elicit many possible hypotheses. We choose these paragraphs to span a range of genres (scientific, biographical, historical, narrative) and present a different set of challenges. For instance, some passages describe physical objects in detail, requiring commonsense understanding of the physical world (e.g., *"... Phonny had not measured his wires in respect to length, but had cut them off of various lengths, taking care however not to have any of them too short. The result was that the ends of the wires projected to various distances above the board... "*). Other passages describe reasoning about uncertainty (e.g., *"... this negative result might be because these animals are not susceptible to the disease... "*) or hypothetical events (e.g., *"... If there should be even partial cooperation between the Austrian leaders, he must retreat ... "*). See Appendix A for the full premise paragraphs. We minimally edit each paragraph so that they can be read standalone, e.g., we resolve coreferences.

3.2 Data Collection Procedure

We collect data over many rounds, where each round comprises three steps. First, crowdworkers write hypothesis sentences that are either entailed or not entailed by one of our premises while interacting with the current model-in-the-loop. Second, other crowdworkers relabel these examples and help filter out spam and other malformed examples. Finally, we update the model-in-the-loop by fine-tuning on all collected data, including data from the newest round. We use Amazon Mechanical Turk (AMT) for data collection.

[1] https://www.gutenberg.org/

Hypothesis Generation. To generate hypotheses, we run AMT tasks where a worker is randomly provided one of the premises and is asked to write ten different hypotheses. After writing each hypothesis, they are shown the predictions of a live model in the loop. To encourage workers to write model-fooling examples, they are given a bonus every time one of their examples fools the model and passes the later label verification step. We ask workers to write ten hypotheses for a single premise, as this allows them to better understand the model's behavior and empirically leads to more-difficult examples (Section 4). The worker can generate hypotheses for either of the binary labels, but we encourage them to generate balanced examples in the onboarding instructions. The user interface is shown in Appendix B.

Label Verification. To ensure the generated hypotheses are labeled correctly, we run a separate AMT task where workers are asked to label each example without being shown the original label. Each example is labeled by at least three workers. If all three agree, that example is saved. If there is a disagreement, we ask two additional workers and keep the example if four out of five agree on the label. We also provide an option to flag a hypothesis as "bad", e.g., it is very ungrammatical or clearly spam. If more than one worker flags an example as bad, we remove it. We do not allow workers to participate in both the labeling and validation AMT tasks, as we do not want workers to be influenced by one another's hypotheses.

Updating the Model. For the initial round of data collection, we use as our starting point a RoBERTa-large model (Liu et al., 2019) that has been finetuned on SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018a), and FEVER-NLI (Nie et al., 2019).[2] We use this training data as it provides us with an accurate initial model, and note that we collapse the neutral and contradiction labels during training as we focus on binary NLI. To update the model after each round, we continue finetuning it on all of the data collected thus far and then deploy it for the next round. Our finetuning hyperparameters follow the recommendations of Mosbach et al. (2021): we use a learning rate of 2×10^{-5}, a learning rate warmup over the first 10% of steps, bias-corrected Adam, and 15 epochs of training. We early stop using held-out validation data (see Section 3.3). We refer to this setting, where crowdworkers interact with a model-in-the-loop that is updated after each round, as the **Dynamic Model** setting.

Baselines. In addition to the above, we also collect data with two baseline approaches:

- **No Model.** This is the typical procedure for collecting training data where workers do not interact with a model.
- **Static Model.** We provide a model in the loop to the workers but the model is kept fixed across all the rounds. We use the same model that the Dynamic Model setting uses in its first round.

No data is mixed between methods and workers can not participate in multiple methods.

3.3 Dataset Details

Our codebase is built on top of the Dynabench platform (Kiela et al., 2021), we deploy tasks using the Mephisto library,[3] and we serve models using Dynalab (Ma et al., 2021). We restrict our AMT workers to those that speak English, have completed at least 100 tasks on AMT, and have an approval rating of at least 97%. To qualify for the task, a worker must also pass an onboarding procedure where they are tasked with correctly labeling five NLI examples in a row.

For each data collection method, we run 20 rounds of data collection. We stop at 20 rounds as model performance on our validation sets begins to saturate. We collect 550 examples per round before label verification, with an equal distribution over the ten premises. All the data collection methods are run in parallel at the same time of day to control for the effects of time on data quality (Karpinska et al., 2021). At this scale, we are able to complete each round of data collection for all three methods in approximately 24 hours. We hold out 50 examples from each round to use for early stopping and for reporting validation metrics.

Table 2 shows overall statistics of our final datasets. These statistics are similar across the three datasets, including the label balance, the rate at which examples are discarded, and the number of AMT workers. However, the datasets differ in the rate at which workers fooled the models in the

[2] FEVER-NLI builds upon the dataset from the FEVER shared task (Thorne et al., 2018). The SNLI and FEVER datasets are licensed CC-BY-SA. MNLI is licensed by MIT. and its copyright is held by New York University (2018).

[3] https://github.com/facebookresearch/mephisto

	No Model	Static Model	Dynamic Model
# Rounds	20	20	20
# Hypo.	11,000	11,000	11,000
# Verified Hypo.	7,684	7,102	6,911
# Workers	115	104	121
% Contradiction	58.5	56.3	54.6

Table 2: *Statistics of our datasets.* For each method, we independently run 20 rounds of data collection with 550 hypotheses per round. We verify the labels of each hypothesis using additional crowd workers and discard any low-agreement examples; the adversarial data is discarded slightly more often. The datasets are roughly balanced between entailment and contradiction.

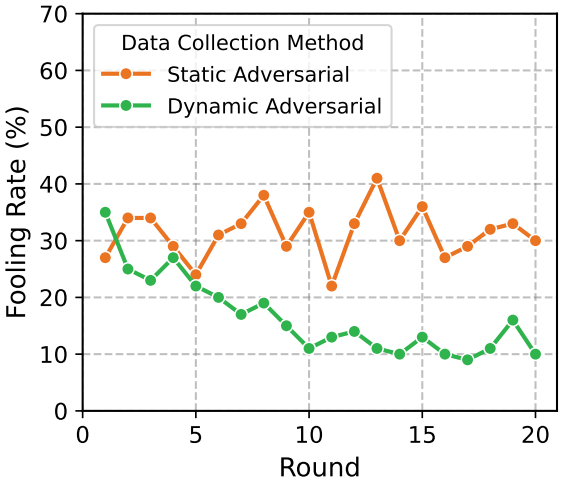

Figure 2: *Model fooling rates.* We show how often crowd workers write examples that are successfully answered by humans but fool the model they interact with. For the static model, the fooling rate is relatively constant as the model is kept fixed (the variance across rounds is due to different crowd workers having different fooling rates). For the dynamic model, the fooling rate goes down over time as the model is updated.

loop; Figure 2 shows that the fooling rate is relatively constant for the static model but goes down for the dynamic model as the model is updated. Table 1 shows qualitative examples of training hypotheses from each method. We will release our data and models publicly.

3.4 Expert-Curated Test Set

Kaushik et al. (2021) compared standard data collection to a single round of adversarial data collection, finding that adversarial training data improves accuracy only on adversarially-constructed test datasets but not on others. We hypothesize that running DADC for many rounds can overcome this limitation and improve generalization to independent, non-adversarial test data. To test this, we built an expert-curated test set for our ten premise paragraphs that is intended to be challenging but not necessarily adversarial to models. We (three of the authors) wrote 680 NLI examples, and we recruited five researchers who have published in NLI and spurious correlations to write an additional 320 examples. The test set spans different challenges, syntactic patterns, and reasoning types, loosely inspired by the categorizations from Williams et al. 2020. The examples are not written with a model in the loop, they are balanced across the labels,[4] and they are equally distributed over the premises. Examples are shown in Table 3.

We also collect crowd worker labels for our test set to ensure that the labels are unambiguous and to measure human accuracy. First, we collect 15 labels for each example. We remove any example from the test set where 9 or fewer workers chose the correct label; this removed 21 examples. Second, we collect an additional 5 labels to use for estimating human accuracy. The average accuracy is **93.2%** when using each label individually.

[4]Our test set is balanced but the three training sets are biased towards contradiction in slightly different amounts. We repeated our experiments by subsampling each training set to be balanced and found nearly identical results.

4 Dynamic Adversarial Data Outperforms Non-Adversarial Data

Here, we show that DADC outperforms both standard and static adversarial data collection in the long term. In particular, we train various models using the three different datasets and compare them on the validation and expert-curated test sets.

4.1 Training Final Models

For each dataset, we train 20 models—one for each round—on all of the training data up to and including a given round. All models start with the same RoBERTa-large model that was used for round one of adversarial data collection. We then continue finetuning this model on the associated training data using the hyperparameters from Section 3.2. Moreover, to measure possible variance across different finetuning runs, we train each model with five different random seeds.

4.2 Main Results

Figure 1 shows our models' accuracy on the expert test set described in Section 3.4. In the short term,

Premise	Hypotheses	Label
Sound	The head of a drum and the strings of a piano are similar in that they both vibrate.	Entailment
	A piano produces sound because the keys vibrate when they are struck by the pianist.	Contradiction
Yellow	The speaker only ran one experiment of injecting yellow fever blood into animals.	Contradiction
	Dr. Daniel Cruz took blood from a sick patient to run his experiment.	Entailment
Faraday	Michael Faraday's wife was named Margaret Hastwell.	Contradiction
	Yorkshire is a less populous locality to be from then Manchester Square.	Entailment

Table 3: Examples from our expert-curated test set. See Table 5 for the premise paragraphs.

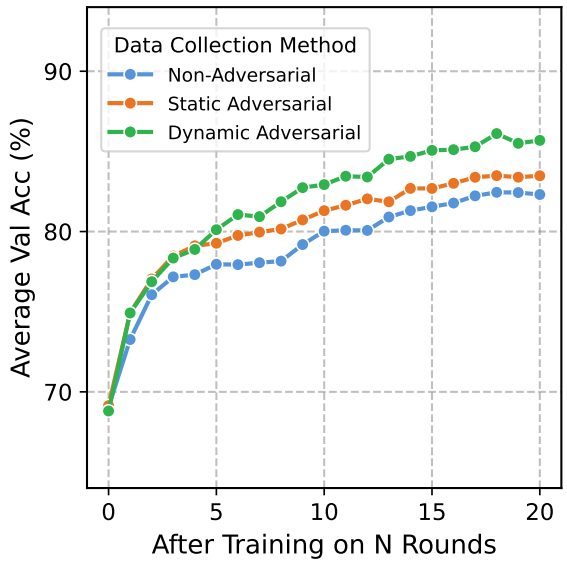

Figure 3: *Combined validation accuracy*. We create a validation set by pooling together validation data from each data collection method. We find the same trend as the expert-curated test set—dynamic adversarial data performs best in the long term.

standard non-adversarial data collection performs best—it has the highest accuracy after the first four rounds. However, in the long term, adversarial data collection, especially when done dynamically, leads to the highest accuracy by a noticeable margin. We run McNemar's statistical test to compute whether the results are significantly different for the final round 20 models: the DADC model outperforms the static adversarial model with $p < 0.05$ and the non-adversarial model with $p < 0.01$; the static adversarial model outperforms the non-adversarial model with $p < 0.05$.

We also evaluate models on validation data that is split off from each round of each data collection method. Figure 3 shows results on a validation set that is created by pooling validation data from all three collection methods; we observe the same trends as our test set, although the accuracies are slightly higher on average.

Overall, these results show that when building training sets in our setting, adversarial data is not necessarily preferred when the number of examples is small. On the contrary, when the number of training examples and rounds is large, using DADC leads to more robust, broader coverage models.

Comparison to Humans. Even though the round 20 models have approximately 700 training examples for each premise, they are still noticeably worse than human accuracy. In particular, the best DADC model reaches 84.4% accuracy, whereas human accuracy is 93.2%. This shows that while DADC does lead to better models, we are still far from creating NLP systems that perform robust NLI on our premise paragraphs.

Generalization of DADC Data Across Models. One possible concern with adversarially-collected data is that it could be too model-specific, similar to datasets built with active learning (Lowell et al., 2019). To test whether the DADC data can generalize to other (newer) models, we train an ALBERT XXLarge-v2 model (Lan et al., 2020) on SNLI, MNLI, and FeverNLI. We then finetune the model on the data from all 20 rounds for each of our three datasets. The model has an accuracy of 69.1% before updating on our collected data, and it reaches an accuracy of 83.1%, 84.6%, and 85.8% on the no model, static model, and dynamic model datasets, respectively. This shows that our DADC data does generalize to better models—it leads to the highest accuracy among the three datasets—but the gap from DADC to static adversarial data is smaller than one from our RoBERTa model.

Generalization Beyond Our Premises. Since the DADC data is more difficult than typical crowdsourced data, it may promote models to learn more robust NLI features. To evaluate this, we test our round 20 models on out-of-distribution datasets, including HANS (McCoy et al., 2019a) and the MNLI mismatched test set (Williams et al., 2018b).

We convert both test sets to binary classification by collapsing the neutral and contradiction labels. We found that the round 20 models from all three settings, as well as our initial model trained on SNLI, MNLI, and FEVER-NLI, reached comparable accuracies on these test sets. This shows that while the DADC data does lead to improved in-distribution test performance, it does not necessarily lead to better performance under distribution shift.

5 Analyzing Adversarial Data

Why is dynamic adversarial data superior to standard data in the long term? In Table 4, we report summary statistics about our three collected datasets. We find that dynamic adversarial data is more diverse, has higher complexity, and contains fewer artifacts than non-adversarial data. These findings agree with our intuition surrounding adversarial datasets: small adversarial training sets that contain diverse and challenging examples may be hard for models to learn from. However, larger datasets of this type will ultimately lead to more accurate and robust models in the long term. We describe our analyses in detail below.[5]

	No Model	Static Model	Dynamic Model
Diversity			
Unique Unigrams	4.0k	4.2k	**4.3k**
Unique Bigrams	23.3k	24.8k	**25.6k**
Inter-example Sim.	41.2	41.9	**39.5**
Complexity			
Syntax	2.0	2.1	**2.3**
Reading Level	4.9	5.4	**5.9**
Length	10.1	10.9	**12.1**
Artifacts			
Hypo-only Acc %	75.4	**69.3**	69.7
Overlap Entail %	54.2	49.2	**47.3**

Table 4: *Dataset analysis.* The hypotheses generated by DADC are more diverse based on the number of lexical items and inter-example similarity scores. The hypotheses are also more complex, as shown by their increased syntactic complexity (Yngve scores), reading level (Flesch-Kincaid readability), and lengths. Finally, adversarial data leads to fewer instances of known artifacts, namely less hypothesis-only information and fewer entailment hypotheses with high lexical overlap. We bold the best result—lower is better for inter-example similarity and the artifact analyses.

Diversity. DADC data is more diverse at both the lexical (unigram and bigram) and example levels (Table 4, top). To measure lexical diversity, we count the number of unique unigrams and bigrams in the dataset. To measure example-level diversity, we iterate through each training example and find the most similar other training sample according to BLEU score (Papineni et al., 2002). We then report the average of these BLEU scores similarities; the dynamic adversarial examples are the least similar to one another.[6] The difference in inter-example similarity between the DADC data and the static adversarial data is significant with $p < 0.01$ according to a t-test.

Syntax and Sentence Complexity. The dynamic adversarial data is more complex (Table 4, middle). For each hypothesis, we measure its length in words, its Flesch-Kincaid readability (Flesch, 1948), and its syntactic complexity using Yngve scores (Yngve, 1960; Roark et al., 2007). Yngve scores roughly measure the deviation of a parse tree from a purely right-branching tree—it is the average number of left branches on the path from the root node to each word. To compute Yngve scores, we parse sentences using the Benepar parser (Kitaev and Klein, 2018) based on T5 small (Raffel et al., 2020). In all three metrics, the dynamic adversarial data scores highest, and it is statistically significantly higher than the static model data based on a t-test with $p < 0.05$. We also show how the syntactic complexity evolves over the rounds in Figure 4. For the non-adversarial and static adversarial data, the syntactic complexity is relatively constant while the DADC examples become increasingly more complex.

Fewer Artifacts. NLI training datasets are known to suffer from spurious correlations. The DADC examples contain fewer instances of two known artifacts: hypothesis-only information (Poliak et al., 2018; Gururangan et al., 2018; Tsuchiya, 2018) and high-overlap entailment examples (McCoy et al., 2019b). To measure such artifacts, we

[5]Note that when computing each metric, we use a version of the No Model and Static Model datasets that are randomly downsampled to be the same size as the dynamic model data (6,911 examples). This controls for any effect that dataset size would have on our analyses.

[6]Note that this diversity metric is effective because we collect hundreds of examples for a single context paragraph; otherwise, we would need to measure similarity between hypotheses for different premises, a more complicated problem. We also experimented with BERTScore (Zhang et al., 2019) and found similar trends as BLEU score.

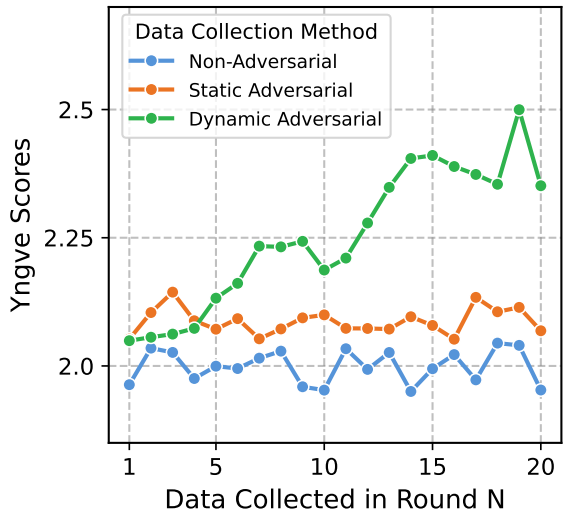

Figure 4: *Complexity of syntax over time.* We show how the average syntactic complexity changes over each round. For the non-adversarial and static adversarial data, the syntactic complexity is relatively constant across rounds. On the other hand, the DADC examples become increasingly more complex as annotators are faced with ever-improving models in the loop.

first train a hypothesis-only model on the training set for each dataset using RoBERTa large. We test on validation data split off from each respective training set, which allows us to measure how much hypothesis-only information is present within each dataset. The static adversarial and dynamic adversarial datasets have the lowest hypothesis-only accuracy. To measure high-overlap entailment instances, we find examples where the hypothesis has high (>90%) word overlap with the premise and compute how often the label is entailment. Such examples appear less frequently in DADC data.

6 Related Work

Post-hoc Adversarial Filtering. In adversarial filtering (Le Bras et al., 2020; Zellers et al., 2018), one takes an existing dataset and trains a model on the most difficult subportion of the data. Adversarial filtering shares motivations with adversarial data collection—difficult examples are more informative for learning—but it is focused on post-hoc data filtering rather than collection of new data. Moreover, in DADC we train on all of the collected examples, whereas adversarial filtering purposefully deletes easy examples.

Active Learning. Active learning (Lewis and Gale, 1994), especially when performed using an uncertainty-based acquisition function, is also closely related to DADC. The key differences are in the setup: in DADC, we need crowdworkers to write novel inputs whereas in active learning one typically assumes access to unlabeled inputs.

Other Data Quality Improvements. Aside from adversarial data collection, researchers have explored numerous methods for improving data quality when using crowdsourcing. This includes feedback from experts (Parrish et al., 2021; Nangia et al., 2021), gamifying the data collection process (Yang et al., 2018), encouraging counterfactual examples (Kaushik et al., 2020; Gardner et al., 2020), or providing prompts that workers can edit (Bowman et al., 2020; Vania et al., 2020). Many of the ideas from these methods can be combined with adversarial data collection, e.g., Eisenschlos et al. (2021) fruitfully combine gamification and adversarial data, and we leave a full exploration of such combinations to future work.

7 Conclusion and Future Work

We investigated dynamic adversarial data collection in the limit—over a large number of rounds until model performance starts plateauing—and demonstrated that data collected via this method is more valuable for training than alternatives, both on validation data and an expert-curated test set. We analyzed the collected data, showing that DADC yields examples that are more diverse, more complex, and contain fewer annotation artifacts compared to non-adversarial examples. Our results show that when building large training sets for training NLP models, data collected in an adversarial fashion with a continually updating model-in-the-loop can be more useful than standard model-agnostic collection in the long term.

In future work, it is vital to conduct similar experiments on different tasks, e.g., question answering and sentiment analysis, as well as on a larger number of contexts for NLI. Such experiments can provide insight into the generalizability of our findings. Moreover, given that a core benefit of DADC is promoting diversity and complexity of examples, one could explore other diversity-promoting methods of data collection. Lastly, our DADC setup is relatively simplistic in that we use a single target model and provide no other guides to the annotator; it would be interesting to provide generative models, model interpretations, an ensemble of target models, or other methods to potentially further improve our DADC results.

Addressing Possible Ethical Concerns

The premises that we use are sourced from publicly available sources and were vetted to ensure they contained no overtly offensive content. As described in main text, we designed our incentive structure to ensure that crowdworkers were well compensated (i.e., paid over minimum wage in the U.S.). Our datasets focus on the English language as it is spoken in the United States. They are not collected for the purpose of designing NLP applications but to conduct a scientific study into collecting data for training machine learning models. We share our datasets to allow the community to replicate our findings and do not foresee any risks associated with the free use of this data.

References

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Rudolf Flesch. 1948. A new readability yardstick. In *Journal of Applied Psychology*.

Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based preselection in crowdsourcing microtasks. In *ACM Transactions of Computer–Human Interaction*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *EMNLP*.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. 2021. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6618–6633, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR*.

Tiffany Liu, A.J. Chavar, Minkyoung Kim, Adam Blufarb, and Rony Karadi. 2021. Sourcing training data with amazon's mechanical turk. *New York Times*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

21–30, Hong Kong, China. Association for Computational Linguistics.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *arXiv preprint arXiv:2106.06052*.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019a. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019b. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, Online. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866. AAAI Press.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *EMNLP Findings*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou,

Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, translational, and clinical language processing*, pages 1–8, Prague, Czech Republic. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Clara Vania, Ruijie Chen, and Samuel R. Bowman. 2020. Asking Crowdworkers to Write Entailment Examples: The Best of Bad options. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 672–686, Suzhou, China. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018a. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018b. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. ANLIzing the adversarial natural language inference dataset. *arXiv preprint arXiv:2010.12729*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Mastering the dungeon: Grounded language learning by mechanical turker descent. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *7th International Conference on Learning Representations, ICLR*.

A Dataset Examples

Table 5 shows the ten paragraphs that are used as the premises in our experiments.

B Mechanical Turk Interface

Figure 5 shows our Amazon Mechanical Turk interface for the model-in-the-loop setting.

Premises
Sound is due to the vibrations of objects. A piano string produces sound because of its vibration when struck, or pulled to one side and then released. This vibration sets the air in rapid motion, and the result is the recording of the sound on our ear-drums. In old telephones, this recording corresponds to a film of sheepskin or bladder drawn over a hollow cup or cylinder. When the head of a drum is struck with a small stick it vibrates. In this case the vibrations are set in motion by the blow, while in the telephone a similar phenomenon is the result of vibratory waves falling from the voice on the thin membrane, or disk of metal, in the transmitter. When these vibrations reach the ear-drumm the nervous system, corresponding to electricity in the mechanical telephone, carries this sound to our brains where it is recorded and understood. In the telephone the wire, charged with electricity, carries the sound from one place to another.
Michael Faraday was born at Newington, Surrey, on September 22, 1791, and was the third of four children. His father, James Faraday, was the son of Robert and Elizabeth Faraday, of Clapham Wood Hall, in the north-west of Yorkshire, and was brought up as a blacksmith. He was the third of ten children, and, in 1786, married Margaret Hastwell, a farmer's daughter. Soon after his marriage he came to London, where Michael was born. In 1796 James Faraday, with his family, moved from Newington, and took rooms over a coach-house in Jacob's Well Mews, Charles Street, Manchester Square. In looking at this humble abode one can scarcely help thinking that the Yorkshire blacksmith and his little family would have been far happier in a country house than in their new crowded London one, however, had he remained in the countryside, it is difficult to see how the genius of young Michael could have met with the requisites for its development.
I had demonstrated by repeated experiments that inoculations of yellow fever blood into animals–dogs, rabbits, guinea pigs–gives a negative result. However, this negative result might be because these animals are not susceptible to the disease. In the civil hospital in Vera Cruz in 1887, Dr. Daniel Ruiz ran a single inoculation experiment on a man. But, this experiment was inconclusive because the patient from whom the blood was obtained was in the eighth day of the disease, and it was quite possible that the specific germ was destoyed at that point. These were the facts surrounding yellow fever when Dr. Reed and his associates commenced their investigations in Cuba during the summer of 1900. In a preliminary note read at the meeting of the American Public Health Association, October 22, 1900, the board gave a report of three cases of yellow fever which they believed to be direct results of mosquito inoculations.
There are other signs of a coming change in the weather known less generally. When birds of long flight, such as swallows and others, hang about home and fly low—rain or wind may be expected. Also when animals seek sheltered places, instead of spreading over their usual range: when pigs carry straw to their sties; and when smoke from chimneys does not ascend readily, an unfavourable change may be looked for. Dew, on the other hand, is an indication of fine weather. So is fog. Neither of of these two formations occurs under an overcast sky, or when there is much wind.
A fierce onslaught was made against Alvinczy's position by Massena's corps. It was entirely unsuccessful, and the French were repulsed with the serious loss of three thousand men. Bonaparte's position was now even more critical than it had been at Castiglione; he had to contend with two new Austrian armies, one on each flank, and Wurmser with a third stood ready to sally out of Mantua in his rear. If there should be even partial cooeperation between the Austrian leaders, he must retreat. But he felt sure there would be no cooeperation whatsoever.
The pendulum had swung—it was no longer the Federalist merchants of New England who were discontent with the policies of the governement, but the planters of the South and particularly of South Carolina. New England was now in favor of a protective tariff. Webster, New England's foremost man at Washington, had voted against the tariff of 1816, but had changed his mind and supported a higher tariff in 1824, and a still higher in 1828. The planters of the South had not found it easy to manufacture goods. They had little or nothing, therefore, to protect against the products of European countries. On the contrary, they exported much to England, and imported from England and other countries many of the things they consumed. Accordingly, they were opposed to the whole system of tariff taxation and desired free trade.
The water was wide and deep, so that he could not cross it. He, however, went down to the brink of the water, and got a good drink. This refreshed him very much, and then he went back again up the bank, and lay down upon the grass there to rest. Presently two cows came down to the water, on the side opposite to where Tony was sitting.
The death of Socrates was brought under three of his enemies—Lycon, Meletus, and Anytus, the last a man of high rank and reputation in the state. Socrates was accused by them of despising the ancient gods of the state, introducing new divinities and corrupting the youth of Athens. He was charged with having taught his followers, young men of the first Athenian families, to despise the established government, to be turbulent and seditious, and his accusors pointed to Alcibiades and Critias, notorious for their lawlessness, as examples of the fruits of his teaching.
In some places the wires came very near together, and in others the spaces between them were so wide, that Wallace thought that the squirrel, if by any chance he should ever get put into the cage, would be very likely to squeeze his way out. Then, besides, Phonny had not measured his wires in respect to length, but had cut them off of various lengths, taking care however not to have any of them too short. The result was that the ends of the wires projected to various distances above the board, presenting a ragged and unworkmanlike appearance.
Garrity was the most sinister figure in organized baseball. Once a newspaper reporter, he had somehow obtained control of the Rockets by chicanery and fraud. Sympathy and gratitude were sentiments unknown to him. He would work a winning pitcher to death, and then send the man shooting down to the minors the moment he showed the slightest symptom of weakness. He scoffed at regulations and bylaws; he defied restraint and control; he was in a constant wrangle with other owners and managers; and as a creator of discord and dissension he held the belt.

Table 5: The ten paragraphs we use as premises in our experiments. We refer to these contexts as Sound, Faraday, Yellow, Weather, Battle, Tariff, Water, Socrates, Wires, and Garrity, respectively.

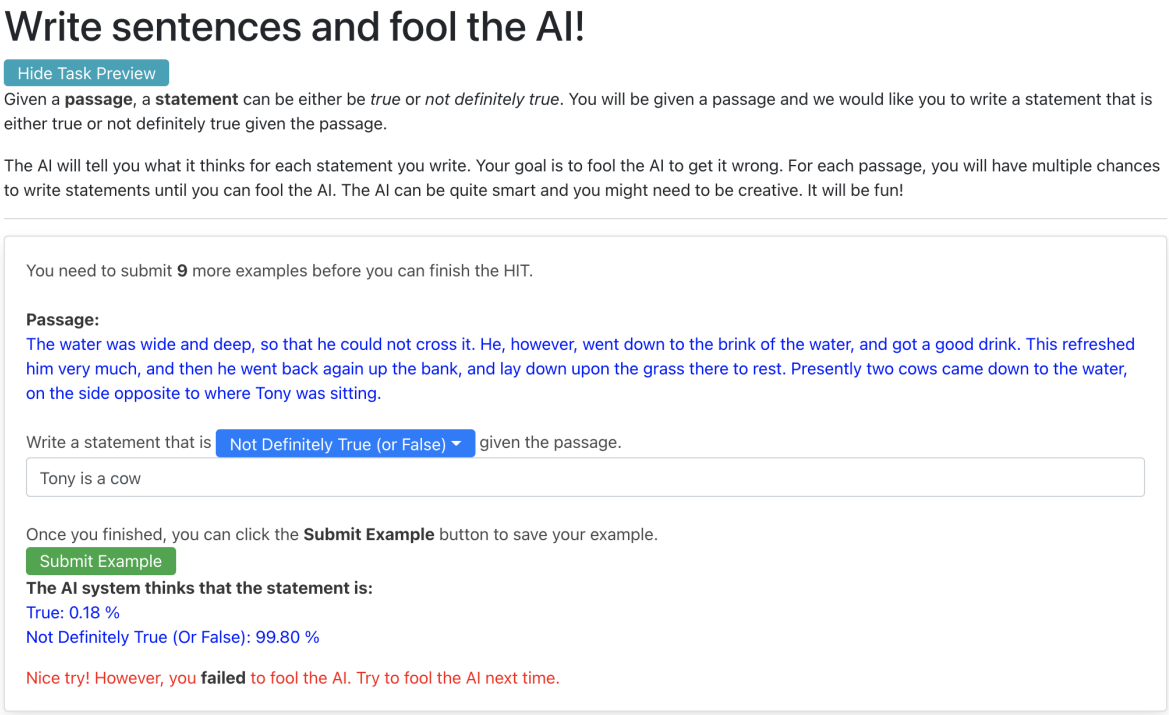

Figure 5: Our Amazon Mechanical Turk interface for the model-in-the-loop setting.