Adversarial NLI: A New Benchmark for Natural Language Understanding

Yixin Nie*, **Adina Williams**†, **Emily Dinan**†, **Mohit Bansal***, **Jason Weston**†, **Douwe Kiela**†
*UNC Chapel Hill
†Facebook AI Research

Abstract

We introduce a new large-scale NLI benchmark dataset, collected via an iterative, adversarial human-and-model-in-the-loop procedure. We show that training models on this new dataset leads to state-of-the-art performance on a variety of popular NLI benchmarks, while posing a more difficult challenge with its new test set. Our analysis sheds light on the shortcomings of current state-of-the-art models, and shows that non-expert annotators are successful at finding their weaknesses. The data collection method can be applied in a never-ending learning scenario, becoming a moving target for NLU, rather than a static benchmark that will quickly saturate.

1 Introduction

Progress in AI has been driven by, among other things, the development of challenging large-scale benchmarks like ImageNet (Russakovsky et al., 2015) in computer vision, and SNLI (Bowman et al., 2015), SQuAD (Rajpurkar et al., 2016), and others in natural language processing (NLP). Recently, for natural language understanding (NLU) in particular, the focus has shifted to combined benchmarks like SentEval (Conneau and Kiela, 2018) and GLUE (Wang et al., 2018), which track model performance on multiple tasks and provide a unified platform for analysis.

With the rapid pace of advancement in AI, however, NLU benchmarks struggle to keep up with model improvement. Whereas it took around 15 years to achieve "near-human performance" on MNIST (LeCun et al., 1998; Cireşan et al., 2012; Wan et al., 2013) and approximately 7 years to surpass humans on ImageNet (Deng et al., 2009; Russakovsky et al., 2015; He et al., 2016), the GLUE benchmark did not last as long as we would have hoped after the advent of BERT (Devlin et al., 2018), and rapidly had to be extended into SuperGLUE (Wang et al., 2019). This raises an important question: Can we collect a large benchmark dataset that can last longer?

The speed with which benchmarks become obsolete raises another important question: are current NLU models genuinely as good as their high performance on benchmarks suggests? A growing body of evidence shows that state-of-the-art models learn to exploit spurious statistical patterns in datasets (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; Glockner et al., 2018; Geva et al., 2019; McCoy et al., 2019), instead of learning *meaning* in the flexible and generalizable way that humans do. Given this, human annotators—be they seasoned NLP researchers or non-experts—might easily be able to construct examples that expose model brittleness.

We propose an iterative, adversarial human-and-model-in-the-loop solution for NLU dataset collection that addresses both benchmark longevity and robustness issues. In the first stage, human annotators devise examples that our current best models cannot determine the correct label for. These resulting hard examples—which should expose additional model weaknesses—can be added to the training set and used to train a stronger model. We then subject the strengthened model to the same procedure and collect weaknesses over several rounds. After each round, we train a new model and set aside a new test set. The process can be iteratively repeated in a never-ending learning (Mitchell et al., 2018) setting, with the model getting stronger and the test set getting harder in each new round. Thus, not only is the resultant dataset harder than existing benchmarks, but this process also yields a "moving post" dynamic target for NLU systems, rather than a static benchmark that will eventually saturate.

Our approach draws inspiration from recent ef-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4885–4901
July 5 - 10, 2020.

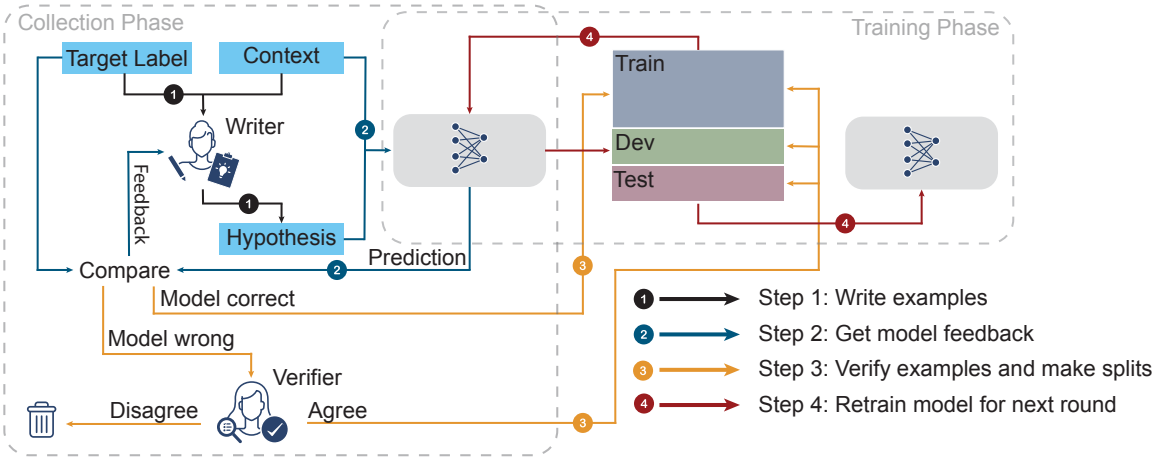

Figure 1: Adversarial NLI data collection via human-and-model-in-the-loop enabled training (HAMLET). The four steps make up one round of data collection. In step 3, model-correct examples are included in the training set; development and test sets are constructed solely from model-wrong verified-correct examples.

forts that gamify collaborative training of machine learning agents over multiple rounds (Yang et al., 2017) and pit "builders" against "breakers" to learn better models (Ettinger et al., 2017). Recently, Dinan et al. (2019) showed that such an approach can be used to make dialogue safety classifiers more robust. Here, we focus on natural language inference (NLI), arguably the most canonical task in NLU. We collected three rounds of data, and call our new dataset Adversarial NLI (ANLI).

Our contributions are as follows: 1) We introduce a novel human-and-model-in-the-loop dataset, consisting of three rounds that progressively increase in difficulty and complexity, that includes annotator-provided explanations. 2) We show that training models on this new dataset leads to state-of-the-art performance on a variety of popular NLI benchmarks. 3) We provide a detailed analysis of the collected data that sheds light on the shortcomings of current models, categorizes the data by inference type to examine weaknesses, and demonstrates good performance on NLI stress tests. The ANLI dataset is available at github.com/facebookresearch/anli/. A demo is available at adversarialnli.com.

2 Dataset collection

The primary aim of this work is to create a new large-scale NLI benchmark on which current state-of-the-art models fail. This constitutes a new target for the field to work towards, and can elucidate model capabilities and limitations. As noted, however, static benchmarks do not last very long these days. If continuously deployed, the data collection procedure we introduce here can pose a dynamic challenge that allows for never-ending learning.

2.1 HAMLET

To paraphrase the great bard (Shakespeare, 1603), *there is something rotten in the state of the art*. We propose *Human-And-Model-in-the-Loop Enabled Training* (HAMLET), a training procedure to automatically mitigate problems with current dataset collection procedures (see Figure 1).

In our setup, our starting point is a *base model*, trained on NLI data. Rather than employing automated adversarial methods, here the model's "adversary" is a human annotator. Given a *context* (also often called a "premise" in NLI), and a desired *target label*, we ask the human *writer* to provide a *hypothesis* that fools the model into misclassifying the label. One can think of the writer as a "white hat" hacker, trying to identify vulnerabilities in the system. For each human-generated example that is misclassified, we also ask the writer to provide a *reason* why they believe it was misclassified.

For examples that the model misclassified, it is necessary to verify that they are actually correct —i.e., that the given context-hypothesis pairs genuinely have their specified target label. The best way to do this is to have them checked by another human. Hence, we provide the example to human *verifiers*. If two human verifiers agree with the writer, the example is considered a good example. If they disagree, we ask a third human verifier to break the tie. If there is still disagreement between the writer and the verifiers, the example is discarded. If the verifiers disagree, they can over-

Context	Hypothesis	Reason	Round	Labels orig.	pred.	valid.	Annotations
Roberto Javier Mora García (c. 1962 – 16 March 2004) was a Mexican journalist and editorial director of "El Mañana", a newspaper based in Nuevo Laredo, Tamaulipas, Mexico. He worked for a number of media outlets in Mexico, including the "El Norte" and "El Diario de Monterrey", prior to his assassination.	Another individual laid waste to Roberto Javier Mora Garcia.	The context states that Roberto Javier Mora Garcia was assassinated, so another person had to have "laid waste to him." The system most likely had a hard time figuring this out due to it not recognizing the phrase "laid waste."	A1 (Wiki)	E	N	E E	Lexical (assassination, laid waste), Tricky (Presupposition), Standard (Idiom)
A melee weapon is any weapon used in direct hand-to-hand combat; by contrast with ranged weapons which act at a distance. The term "melee" originates in the 1640s from the French word "mêlée", which refers to hand-to-hand combat, a close quarters battle, a brawl, a confused fight, etc. Melee weapons can be broadly divided into three categories	Melee weapons are good for ranged and hand-to-hand combat.	Melee weapons are good for hand to hand combat, but NOT ranged.	A2 (Wiki)	C	E	C N C	Standard (Conjunction), Tricky (Exhaustification), Reasoning (Facts)
If you can dream it, you can achieve it—unless you're a goose trying to play a very human game of rugby. In the video above, one bold bird took a chance when it ran onto a rugby field mid-play. Things got dicey when it got into a tussle with another player, but it shook it off and kept right on running. After the play ended, the players escorted the feisty goose off the pitch. It was a risky move, but the crowd chanting its name was well worth it.	The crowd believed they knew the name of the goose running on the field.	Because the crowd was chanting its name, the crowd must have believed they knew the goose's name. The word "believe" may have made the system think this was an ambiguous statement.	A3 (News)	E	N	E E	Reasoning (Facts), Reference (Coreference)

Table 1: Examples from development set. 'An' refers to round number, 'orig.' is the original annotator's gold label, 'pred.' is the model prediction, 'valid.' are the validator labels, 'reason' was provided by the original annotator, 'Annotations' are the tags determined by an linguist expert annotator.

rule the original target label of the writer.

Once data collection for the current round is finished, we construct a new training set from the collected data, with accompanying development and test sets, which are constructed solely from verified correct examples. The test set was further restricted so as to: 1) include pairs from "exclusive" annotators who are never included in the training data; and 2) be balanced by label classes (and genres, where applicable). We subsequently train a *new model* on this and other existing data, and repeat the procedure.

2.2 Annotation details

We employed Mechanical Turk workers with qualifications and collected hypotheses via the ParlAI[1] framework. Annotators are presented with a context and a target label—either 'entailment', 'contradiction', or 'neutral'—and asked to write a hypothesis that corresponds to the label. We phrase the label classes as "definitely correct", "definitely incorrect", or "neither definitely correct nor definitely incorrect" given the context, to make the task easier to grasp. Model predictions are obtained for the context and submitted hypothesis pair. The probability of each label is shown to the worker as feedback. If the model prediction was incorrect, the job is complete. If not, the worker continues to write hypotheses for the given (context, target-label) pair until the model predicts the label incorrectly or the number of tries exceeds a threshold (5 tries in the first round, 10 tries thereafter).

To encourage workers, payments increased as rounds became harder. For hypotheses that the model predicted incorrectly, and that were verified by other humans, we paid an additional bonus on top of the standard rate.

2.3 Round 1

For the first round, we used a BERT-Large model (Devlin et al., 2018) trained on a concatenation of SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2017), and selected the best-performing model we could train as the starting point for our dataset collection procedure. For Round 1 contexts, we randomly sampled short multi-sentence passages from Wikipedia (of 250-600 characters) from the manually curated HotpotQA training set (Yang et al., 2018). Contexts are either ground-truth contexts from that dataset, or they are Wikipedia passages retrieved using TF-IDF (Chen et al., 2017) based on a HotpotQA question.

2.4 Round 2

For the second round, we used a more powerful RoBERTa model (Liu et al., 2019b) trained on SNLI, MNLI, an NLI-version[2] of FEVER (Thorne et al., 2018), and the training data from the previous round (A1). After a hyperparameter search, we

[1] https://parl.ai/

[2] The NLI version of FEVER pairs claims with evidence retrieved by Nie et al. (2019) as (context, hypothesis) inputs.

Dataset	Genre	Context	Train / Dev / Test	Model error rate		Tries	Time (sec.)
				Unverified	Verified	mean/median per verified ex.	
A1	Wiki	2,080	16,946 / 1,000 / 1,000	29.68%	18.33%	3.4 / 2.0	199.2 / 125.2
A2	Wiki	2,694	45,460 / 1,000 / 1,000	16.59%	8.07%	6.4 / 4.0	355.3 / 189.1
A3	Various	6,002	100,459 / 1,200 / 1,200	17.47%	8.60%	6.4 / 4.0	284.0 / 157.0
	(Wiki subset)	1,000	19,920 / 200 / 200	14.79%	6.92%	7.4 / 5.0	337.3 / 189.6
ANLI	Various	10,776	162,865 / 3,200 / 3,200	18.54%	9.52%	5.7 / 3.0	282.9 / 156.3

Table 2: Dataset statistics: 'Model error rate' is the percentage of examples that the model got wrong; 'unverified' is the overall percentage, while 'verified' is the percentage that was verified by at least 2 human annotators.

selected the model with the best performance on the A1 development set. Then, using the hyperparameters selected from this search, we created a final set of models by training several models with different random seeds. During annotation, we constructed an ensemble by randomly picking a model from the model set as the adversary each turn. This helps us avoid annotators exploiting vulnerabilities in one single model. A new non-overlapping set of contexts was again constructed from Wikipedia via HotpotQA using the same method as Round 1.

2.5 Round 3

For the third round, we selected a more diverse set of contexts, in order to explore robustness under domain transfer. In addition to contexts from Wikipedia for Round 3, we also included contexts from the following domains: News (extracted from Common Crawl), fiction (extracted from StoryCloze (Mostafazadeh et al., 2016) and CBT (Hill et al., 2015)), formal spoken text (excerpted from court and presidential debate transcripts in the Manually Annotated Sub-Corpus (MASC) of the Open American National Corpus[3]), and causal or procedural text, which describes sequences of events or actions, extracted from WikiHow. Finally, we also collected annotations using the longer contexts present in the GLUE RTE training data, which came from the RTE5 dataset (Bentivogli et al., 2009). We trained an even stronger RoBERTa ensemble by adding the training set from the second round (A2) to the training data.

2.6 Comparing with other datasets

The ANLI dataset, comprising three rounds, improves upon previous work in several ways. First, and most obviously, the dataset is collected to be more difficult than previous datasets, by design. Second, it remedies a problem with SNLI, namely that its contexts (or premises) are very short, because they were selected from the image captioning domain. We believe longer contexts should naturally lead to harder examples, and so we constructed ANLI contexts from longer, multi-sentence source material.

Following previous observations that models might exploit spurious biases in NLI hypotheses, (Gururangan et al., 2018; Poliak et al., 2018), we conduct a study of the performance of hypothesis-only models on our dataset. We show that such models perform poorly on our test sets.

With respect to data generation with naïve annotators, Geva et al. (2019) noted that models can pick up on annotator bias, modelling annotator artefacts rather than the intended reasoning phenomenon. To counter this, we selected a subset of annotators (i.e., the "exclusive" workers) whose data would only be included in the test set. This enables us to avoid overfitting to the writing style biases of particular annotators, and also to determine how much individual annotator bias is present for the main portion of the data. Examples from each round of dataset collection are provided in Table 1.

Furthermore, our dataset poses new challenges to the community that were less relevant for previous work, such as: can we improve performance online without having to train a new model from scratch every round, how can we overcome catastrophic forgetting, how do we deal with mixed model biases, etc. Because the training set includes examples that the model got right but were not verified, learning from noisy and potentially unverified data becomes an additional interesting challenge.

3 Dataset statistics

The dataset statistics can be found in Table 2. The number of examples we collected increases per round, starting with approximately 19k examples for Round 1, to around 47k examples for Round 2,

[3] anc.org/data/masc/corpus/

Model	Training Data	A1	A2	A3	ANLI	ANLI-E	SNLI	MNLI-m/-mm
BERT	S,M$^{\star 1}$	00.0	28.9	28.8	19.8	19.9	91.3	86.7 / 86.4
	+A1	44.2	32.6	29.3	35.0	34.2	91.3	86.3 / 86.5
	+A1+A2	57.3	45.2	33.4	44.6	43.2	90.9	86.3 / 86.3
	+A1+A2+A3	57.2	49.0	46.1	50.5	46.3	90.9	85.6 / 85.4
	S,M,F,ANLI	57.4	48.3	43.5	49.3	44.2	90.4	86.0 / 85.8
XLNet	S,M,F,ANLI	67.6	50.7	48.3	55.1	52.0	91.8	89.6 / 89.4
RoBERTa	S,M	47.6	25.4	22.1	31.1	31.4	92.6	90.8 / 90.6
	+F	54.0	24.2	22.4	32.8	33.7	92.7	90.6 / 90.5
	+F+A1$^{\star 2}$	68.7	19.3	22.0	35.8	36.8	92.8	90.9 / 90.7
	+F+A1+A2$^{\star 3}$	71.2	44.3	20.4	43.7	41.4	92.9	91.0 / 90.7
	S,M,F,ANLI	73.8	48.9	44.4	53.7	49.7	92.6	91.0 / 90.6

Table 3: Model Performance. 'S' refers to SNLI, 'M' to MNLI dev (-m=matched, -mm=mismatched), and 'F' to FEVER; 'A1–A3' refer to the rounds respectively and 'ANLI' refers to A1+A2+A3, '-E' refers to test set examples written by annotators exclusive to the test set. Datasets marked '$\star^n$' were used to train the base model for round n, and their performance on that round is underlined (A2 and A3 used ensembles, and hence have non-zero scores).

to over 103k examples for Round 3. We collected more data for later rounds not only because that data is likely to be more interesting, but also simply because the base model is better and so annotation took longer to collect good, verified correct examples of model vulnerabilities.

For each round, we report the model error rate, both on verified and unverified examples. The unverified model error rate captures the percentage of examples where the model disagreed with the writer's target label, but where we are not (yet) sure if the example is correct. The verified model error rate is the percentage of model errors from example pairs that other annotators confirmed the correct label for. Note that error rate is a useful way to evaluate model quality: the lower the model error rate—assuming constant annotator quality and context-difficulty—the better the model.

We observe that model error rates decrease as we progress through rounds. In Round 3, where we included a more diverse range of contexts from various domains, the overall error rate went slightly up compared to the preceding round, but for Wikipedia contexts the error rate decreased substantially. While for the first round roughly 1 in every 5 examples were verified model errors, this quickly dropped over consecutive rounds, and the overall model error rate is less than 1 in 10. On the one hand, this is impressive, and shows how far we have come with just three rounds. On the other hand, it shows that we still have a long way to go if even untrained annotators can fool ensembles of state-of-the-art models with relative ease.

Table 2 also reports the average number of "tries", i.e., attempts made for each context until a model error was found (or the number of possible tries is exceeded), and the average time this took (in seconds). Again, these metrics are useful for evaluating model quality: observe that the average number of tries and average time per verified error both go up with later rounds. This demonstrates that the rounds are getting increasingly more difficult. Further dataset statistics and inter-annotator agreement are reported in Appendix C.

4 Results

Table 3 reports the main results. In addition to BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019b), we also include XLNet (Yang et al., 2019) as an example of a strong, but different, model architecture. We show test set performance on the ANLI test sets per round, the total ANLI test set, and the exclusive test subset (examples from test-set-exclusive workers). We also show accuracy on the SNLI test set and the MNLI development set (for the purpose of comparing between different model configurations across table rows). In what follows, we discuss our observations.

Base model performance is low. Notice that the base model for each round performs very poorly on that round's test set. This is the expected outcome: For round 1, the base model gets the entire test set wrong, by design. For rounds 2 and 3, we used an ensemble, so performance is not necessarily zero. However, as it turns out, performance still falls well below chance[4], indicating that workers did not find vulnerabilities specific to a single model, but generally applicable ones for that model class.

[4]Chance is at 33%, since the test set labels are balanced.

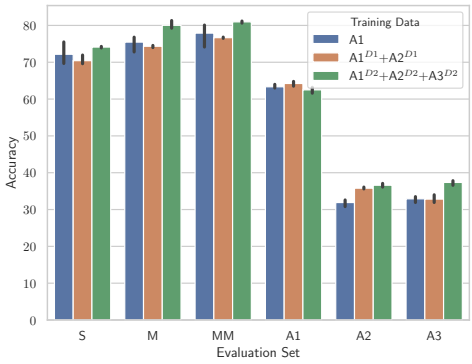

Figure 2: RoBERTa performance on dev, with A1–3 downsampled s.t. $|A1^{D1}|=|A2^{D1}|=\frac{1}{2}|A1|$ and $|A1^{D2}|=|A2^{D2}|=|A3^{D2}|=\frac{1}{3}|A1|$.

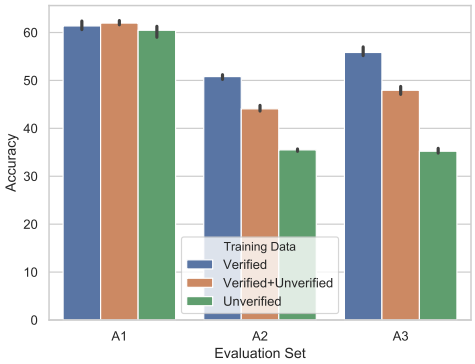

Figure 3: Comparison of verified, unverified and combined data, where data sets are downsampled to ensure equal training sizes.

Rounds become increasingly more difficult. As already foreshadowed by the dataset statistics, round 3 is more difficult (yields lower performance) than round 2, and round 2 is more difficult than round 1. This is true for all model architectures.

Training on more rounds improves robustness. Generally, our results indicate that training on more rounds improves model performance. This is true for all model architectures. Simply training on more "normal NLI" data would not help a model be robust to adversarial attacks, but our data actively helps mitigate these.

RoBERTa achieves state-of-the-art performance... We obtain state of the art performance on both SNLI and MNLI with the RoBERTa model finetuned on our new data. The RoBERTa paper (Liu et al., 2019b) reports a score of 90.2 for both MNLI-matched and -mismatched dev, while we obtain 91.0 and 90.7. The state of the art on SNLI is currently held by MT-DNN (Liu et al., 2019a), which reports 91.6 compared to our 92.9.

...but is outperformed when it is base model. However, the base (RoBERTa) models for rounds 2 and 3 are outperformed by both BERT and XLNet (rows 5, 6 and 10). This shows that annotators found examples that RoBERTa generally struggles with, which cannot be mitigated by more examples alone. It also implies that BERT, XLNet, and RoBERTa all have different weaknesses, possibly as a function of their training data (BERT, XLNet and RoBERTa were trained on different data sets, which might or might not have contained information relevant to the weaknesses).

Continuously augmenting training data does not downgrade performance. Even though ANLI training data is different from SNLI and MNLI, adding it to the training set does not harm performance on those tasks. Our results (see also rows 2-3 of Table 6) suggest the method could successfully be applied for multiple additional rounds.

Exclusive test subset difference is small. We included an exclusive test subset (ANLI-E) with examples from annotators never seen in training, and find negligible differences, indicating that our models do not over-rely on annotator's writing styles.

4.1 The effectiveness of adversarial training

We examine the effectiveness of the adversarial training data in two ways. First, we sample from respective datasets to ensure exactly equal amounts of training data. Table 5 shows that the adversarial data improves performance, including on SNLI and MNLI when we replace part of those datasets with the adversarial data. This suggests that the adversarial data is more data efficient than "normally collected" data. Figure 2 shows that adversarial data collected in later rounds is of higher quality and more data-efficient.

Second, we compared verified correct examples of model vulnerabilities (examples that the model got wrong and were verified to be correct) to unverified ones. Figure 3 shows that the verified correct examples are much more valuable than the unverified examples, especially in the later rounds (where the latter drops to random).

4.2 Stress Test Results

We also test models on two recent hard NLI test sets: SNLI-Hard (Gururangan et al., 2018) and

Model	SNLI-Hard	NLI Stress Tests					
		AT (m/mm)	NR	LN (m/mm)	NG (m/mm)	WO (m/mm)	SE (m/mm)
Previous models	72.7	14.4 / 10.2	28.8	58.7 / 59.4	48.8 / 46.6	50.0 / 50.2	58.3 / 59.4
BERT (All)	82.3	75.0 / 72.9	65.8	84.2 / 84.6	64.9 / 64.4	61.6 / 60.6	78.3 / 78.3
XLNet (All)	83.5	88.2 / 87.1	85.4	87.5 / 87.5	59.9 / 60.0	68.7 / 66.1	84.3 / 84.4
RoBERTa (S+M+F)	84.5	81.6 / 77.2	62.1	88.0 / 88.5	61.9 / 61.9	67.9 / 66.2	86.2 / 86.5
RoBERTa (All)	84.7	85.9 / 82.1	80.6	88.4 / 88.5	62.2 / 61.9	67.4 / 65.6	86.3 / 86.7

Table 4: Model Performance on NLI stress tests (tuned on their respective dev. sets). All=S+M+F+ANLI. AT='Antonym'; 'NR'=Numerical Reasoning; 'LN'=Length; 'NG'=Negation; 'WO'=Word Overlap; 'SE'=Spell Error. Previous models refers to the Naik et al. (2018) implementation of Conneau et al. (2017, InferSent) for the Stress Tests, and to the Gururangan et al. (2018) implementation of Gong et al. (2018, DIIN) for SNLI-Hard.

Train Data	A1	A2	A3	S	M-m/mm
SM^{D1}+SM^{D2}	45.1	26.1	27.1	**92.5**	89.8/**89.7**
SM^{D1}+A	**72.6**	**42.9**	**42.0**	92.3	**90.3**/89.6
SM	48.0	24.8	31.1	93.2	90.8/90.6
SM^{D3}+A	**73.3**	**42.4**	**40.5**	**93.3**	**90.8/90.7**

Table 5: RoBERTa performance on dev set with different training data. S=SNLI, M=MNLI, A=A1+A2+A3. 'SM' refers to combined S and M training set. D1, D2, D3 means down-sampling SM s.t. $|SM^{D2}|=|A|$ and $|SM^{D3}|+|A|=|SM|$. Therefore, training sizes are identical in every pair of rows.

Train Data	A1	A2	A3	S	M-m/mm
ALL	73.8	48.9	44.4	92.6	91.0/90.6
S+M	47.6	25.4	22.1	92.6	90.8/90.6
ANLI-Only	71.3	43.3	43.0	83.5	86.3/86.5
ALL^{H}	49.7	46.3	42.8	71.4	60.2/59.8
$S+M^{H}$	33.1	29.4	32.2	71.8	62.0/62.0
$ANLI\text{-}Only^{H}$	51.0	42.6	41.5	47.0	51.9/54.5

Table 6: Performance of RoBERTa with different data combinations. ALL=S,M,F,ANLI. Hypothesis-only models are marked H where they are trained and tested with only hypothesis texts.

the NLI stress tests (Naik et al., 2018) (see Appendix A for details). The results are in Table 4. We observe that all our models outperform the models presented in original papers for these common stress tests. The RoBERTa models perform best on SNLI-Hard and achieve accuracy levels in the high 80s on the 'antonym' (AT), 'numerical reasoning' (NR), 'length' (LN), 'spelling error'(SE) sub-datasets, and show marked improvement on both 'negation' (NG), and 'word overlap' (WO). Training on ANLI appears to be particularly useful for the AT, NR, NG and WO stress tests.

4.3 Hypothesis-only results

For SNLI and MNLI, concerns have been raised about the propensity of models to pick up on spurious artifacts that are present just in the hypotheses (Gururangan et al., 2018; Poliak et al., 2018). Here, we compare full models to models trained only on the hypothesis (marked H). Table 6 reports results on ANLI, as well as on SNLI and MNLI. The table shows that hypothesis-only models perform poorly on ANLI[5], and obtain good performance on SNLI and MNLI. Hypothesis-only performance decreases over rounds for ANLI.

We observe that in rounds 2 and 3, RoBERTa is not much better than hypothesis-only. This could mean two things: either the test data is very difficult, or the training data is not good. To rule out the latter, we trained only on ANLI ($\sim$163k training examples): RoBERTa matches BERT when trained on the much larger, fully in-domain SNLI+MNLI combined dataset (943k training examples) on MNLI, with both getting $\sim$86 (the third row in Table 6). Hence, this shows that the test sets are so difficult that state-of-the-art models cannot outperform a hypothesis-only prior.

5 Linguistic analysis

We explore the types of inferences that fooled models by manually annotating 500 examples from each round's development set. A dynamically evolving dataset offers the unique opportunity to track how model error rates change over time. Since each round's development set contains only verified examples, we can investigate two interesting questions: which types of inference do writers employ to fool the models, and are base models differentially sensitive to different types of reasoning?

The results are summarized in Table 7. We devised an inference ontology containing six types of inference: Numerical & Quantitative (i.e., reason-

[5]Obviously, without manual intervention, some bias remains in how people phrase hypotheses—e.g., contradiction might have more negation—which explains why hypothesis-only performs slightly above chance when trained on ANLI.

Round	Numerical & Quant.	Reference & Names	Standard	Lexical	Tricky	Reasoning & Facts	Quality
A1	38%	13%	18%	13%	22%	53%	4%
A2	32%	20%	21%	21%	20%	59%	3%
A3	10%	18%	27%	27%	27%	63%	3%
Average	27%	17%	22%	22%	23%	58%	3%

Table 7: Analysis of 500 development set examples per round and on average.

ing about cardinal and ordinal numbers, inferring dates and ages from numbers, etc.), Reference & Names (coreferences between pronouns and forms of proper names, knowing facts about name gender, etc.), Standard Inferences (conjunctions, negations, cause-and-effect, comparatives and superlatives etc.), Lexical Inference (inferences made possible by lexical information about synonyms, antonyms, etc.), Tricky Inferences (wordplay, linguistic strategies such as syntactic transformations/reorderings, or inferring writer intentions from contexts), and reasoning from outside knowledge or additional facts (e.g., "You can't reach the sea directly from Rwanda"). The quality of annotations was also tracked; if a pair was ambiguous or a label debatable (from the expert annotator's perspective), it was flagged. Quality issues were rare at 3-4% per round. Any one example can have multiple types, and every example had at least one tag.

We observe that both round 1 and 2 writers rely heavily on numerical and quantitative reasoning in over 30% of the development set—the percentage in A2 (32%) dropped roughly 6% from A1 (38%)—while round 3 writers use numerical or quantitative reasoning for only 17%. The majority of numerical reasoning types were references to cardinal numbers that referred to dates and ages. Inferences predicated on references and names were present in about 10% of rounds 1 & 3 development sets, and reached a high of 20% in round 2, with coreference featuring prominently. Standard inference types increased in prevalence as the rounds increased, ranging from 18%–27%, as did 'Lexical' inferences (increasing from 13%–31%). The percentage of sentences relying on reasoning and outside facts remains roughly the same, in the mid-50s, perhaps slightly increasing over the rounds. For round 3, we observe that the model used to collect it appears to be more susceptible to Standard, Lexical, and Tricky inference types. This finding is compatible with the idea that models trained on adversarial data perform better, since annotators seem to have been encouraged to devise more creative examples containing harder types of inference in order to stump them. Further analysis is provided in Appendix B.

6 Related work

Bias in datasets Machine learning methods are well-known to pick up on spurious statistical patterns. For instance, in the first visual question answering dataset (Antol et al., 2015), biases like "2" being the correct answer to 39% of the questions starting with "how many" allowed learning algorithms to perform well while ignoring the visual modality altogether (Jabri et al., 2016; Goyal et al., 2017). In NLI, Gururangan et al. (2018), Poliak et al. (2018) and Tsuchiya (2018) showed that hypothesis-only baselines often perform far better than chance. NLI systems can often be broken merely by performing simple lexical substitutions (Glockner et al., 2018), and struggle with quantifiers (Geiger et al., 2018) and certain superficial syntactic properties (McCoy et al., 2019).

In question answering, Kaushik and Lipton (2018) showed that question- and passage-only models can perform surprisingly well, while Jia and Liang (2017) added adversarially constructed sentences to passages to cause a drastic drop in performance. Many tasks do not actually require sophisticated linguistic reasoning, as shown by the surprisingly good performance of random encoders (Wieting and Kiela, 2019). Similar observations were made in machine translation (Belinkov and Bisk, 2017) and dialogue (Sankar et al., 2019). Machine learning also has a tendency to overfit on static targets, even if that does not happen deliberately (Recht et al., 2018). In short, the field is rife with dataset bias and papers trying to address this important problem. This work presents a potential solution: if such biases exist, they will allow humans to fool the models, resulting in valuable training examples until the bias is mitigated.

Dynamic datasets. Bras et al. (2020) proposed AFLite, an approach for avoiding spurious biases through adversarial filtering, which is a model-in-the-loop approach that iteratively probes and improves models. Kaushik et al. (2019) offer a

causal account of spurious patterns, and counterfactually augment NLI datasets by editing examples to break the model. That approach is human-in-the-loop, using humans to find problems with one single model. In this work, we employ both human and model-based strategies iteratively, in a form of human-and-model-in-the-loop training, to create completely *new* examples, in a potentially never-ending loop (Mitchell et al., 2018).

Human-and-model-in-the-loop training is not a new idea. Mechanical Turker Descent proposes a gamified environment for the collaborative training of grounded language learning agents over multiple rounds (Yang et al., 2017). The "Build it Break it Fix it" strategy in the security domain (Ruef et al., 2016) has been adapted to NLP (Ettinger et al., 2017) as well as dialogue safety (Dinan et al., 2019). The QApedia framework (Kratzwald and Feuerriegel, 2019) continuously refines and updates its content repository using humans in the loop, while human feedback loops have been used to improve image captioning systems (Ling and Fidler, 2017). Wallace et al. (2019) leverage trivia experts to create a model-driven adversarial question writing procedure and generate a small set of challenge questions that QA-models fail on. Relatedly, Lan et al. (2017) propose a method for continuously growing a dataset of paraphrases.

There has been a flurry of work in constructing datasets with an adversarial component, such as Swag (Zellers et al., 2018) and HellaSwag (Zellers et al., 2019), CODAH (Chen et al., 2019), Adversarial SQuAD (Jia and Liang, 2017), Lambada (Paperno et al., 2016) and others. Our dataset is not to be confused with abductive NLI (Bhagavatula et al., 2019), which calls itself αNLI, or ART.

7 Discussion & Conclusion

In this work, we used a human-and-model-in-the-loop training method to collect a new benchmark for natural language understanding. The benchmark is designed to be challenging to current state-of-the-art models. Annotators were employed to act as adversaries, and encouraged to find vulnerabilities that fool the model into misclassifying, but that another person would correctly classify. We found that non-expert annotators, in this gamified setting and with appropriate incentives, are remarkably creative at finding and exploiting weaknesses. We collected three rounds, and as the rounds progressed, the models became more robust and the test sets for each round became more difficult. Training on this new data yielded the state of the art on existing NLI benchmarks.

The ANLI benchmark presents a new challenge to the community. It was carefully constructed to mitigate issues with previous datasets, and was designed from first principles to last longer. The dataset also presents many opportunities for further study. For instance, we collected annotator-provided explanations for each example that the model got wrong. We provided inference labels for the development set, opening up possibilities for interesting more fine-grained studies of NLI model performance. While we verified the development and test examples, we did not verify the correctness of each training example, which means there is probably some room for improvement there.

A concern might be that the static approach is probably cheaper, since dynamic adversarial data collection requires a verification step to ensure examples are correct. However, verifying examples is probably also a good idea in the static case, and adversarially collected examples can still prove useful even if they didn't fool the model and weren't verified. Moreover, annotators were better incentivized to do a good job in the adversarial setting. Our finding that adversarial data is more data-efficient corroborates this theory. Future work could explore a detailed cost and time trade-off between adversarial and static collection.

It is important to note that our approach is model-agnostic. HAMLET was applied against an ensemble of models in rounds 2 and 3, and it would be straightforward to put more diverse ensembles in the loop to examine what happens when annotators are confronted with a wider variety of architectures.

The proposed procedure can be extended to other classification tasks, as well as to ranking with hard negatives either generated (by adversarial models) or retrieved and verified by humans. It is less clear how the method can be applied in generative cases.

Adversarial NLI is meant to be a challenge for measuring NLU progress, even for as yet undiscovered models and architectures. Luckily, if the benchmark does turn out to saturate quickly, we will always be able to collect a new round.

Acknowledgments

YN interned at Facebook. YN and MB were sponsored by DARPA MCS Grant #N66001-19-2-4031, ONR Grant #N00014-18-1-2871, and DARPA YFA17-D17AP00022. Special thanks to Sam Bowman for comments on an earlier draft.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. *TAC*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *arXiv preprint arXiv:2002.04108*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: an adversarially authored question-answer dataset for common sense. *CoRR*, abs/1904.04365.

Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. 2012. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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 EMNLP*.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M Bender. 2017. Towards linguistically generalizable nlp systems: A workshop and shared task. *arXiv preprint arXiv:1711.01505*.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2018. Stress-testing neural models of natural language inference with multiply-quantified sentences. *arXiv preprint arXiv:1810.13033*.

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. *arXiv preprint arXiv:1908.07898*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of ACL*.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of EMNLP*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Bernhard Kratzwald and Stefan Feuerriegel. 2019. Learning from on-line user feedback in neural question answering on the web. In *The World Wide Web Conference*, pages 906–916. ACM.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Huan Ling and Sanja Fidler. 2017. Teaching machines to describe images via natural language feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5075–5085.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

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

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. 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.

Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, B Dalvi, Matt Gardner, Bryan Kisiel, et al. 2018. Never-ending learning. *Communications of the ACM*, 61(5):103–115.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

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.

Yixin Nie and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. *arXiv preprint arXiv:1708.02312*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

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*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2018. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*.

Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L Mazurek, and Piotr Mardziel. 2016. Build it, break it, fix it: Contesting secure development. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 690–703. ACM.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. *arXiv preprint arXiv:1906.01603*.

William Shakespeare. 1603. *The Tragedy of Hamlet, Prince of Denmark*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of LREC*.

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 question answering examples. In *Transactions of the Association for Computational Linguistics*.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. 2013. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. *arXiv preprint arXiv:1901.10444*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. 2017. Mastering the dungeon: Grounded language learning by mechanical turker descent. *arXiv preprint arXiv:1711.07950*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of EMNLP*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of ACL*.

A Performance on challenge datasets

Recently, several hard test sets have been made available for revealing the biases NLI models learn from their training datasets (Nie and Bansal, 2017; McCoy et al., 2019; Gururangan et al., 2018; Naik et al., 2018). We examine model performance on two of these: the SNLI-Hard (Gururangan et al., 2018) test set, which consists of examples that hypothesis-only models label incorrectly, and the NLI stress tests (Naik et al., 2018), in which sentences containing antonyms pairs, negations, high word overlap, i.a., are heuristically constructed. We test our models on these stress tests after tuning on each test's respective development set to account for potential domain mismatches. For comparison, we also report results from the original papers: for SNLI-Hard from Gururangan et al.'s implementation of the hierarchical tensor-based Densely Interactive Inference Network (Gong et al., 2018, DIIN) on MNLI, and for the NLI stress tests, Naik et al.'s implementation of InferSent (Conneau et al., 2017) trained on SNLI.

B Further linguistic analysis

We compare the incidence of linguistic phenomena in ANLI with extant popular NLI datasets to get an idea of what our dataset contains. We observe that FEVER and SNLI datasets generally contain many fewer hard linguistic phenomena than MultiNLI and ANLI (see Table 8).

ANLI and MultiNLI have roughly the same percentage of hypotheses that exceeding twenty words in length, and/or contain negation (e.g., 'never', 'no'), tokens of 'or', and modals (e.g., 'must', 'can'). MultiNLI hypotheses generally contains more pronouns, quantifiers (e.g., 'many', 'every'), WH-words (e.g., 'who', 'why'), and tokens of 'and' than do their ANLI counterparts—although A3 reaches nearly the same percentage as MultiNLI for negation, and modals. However, ANLI contains more cardinal numerals and time terms (such as 'before', 'month', and 'tomorrow') than MultiNLI. These differences might be due to the fact that the two datasets are constructed from different genres of text. Since A1 and A2 contexts are constructed from a single Wikipedia data source (i.e., HotPotQA data), and most Wikipedia articles include dates in the first line, annotators appear to prefer constructing hypotheses that highlight numerals and time terms, leading to their high incidence.

Focusing on ANLI more specifically, A1 has roughly the same incidence of most tags as A2 (i.e., within 2% of each other), which, again, accords with the fact that we used the same Wikipedia data source for A1 and A2 contexts. A3, however, has the highest incidence of every tag (except for numbers and time) in the ANLI dataset. This could be due to our sampling of A3 contexts from a wider range of genres, which likely affected how annotators chose to construct A3 hypotheses; this idea is supported by the fact that A3 contexts differ in tag percentage from A1 and A2 contexts as well. The higher incidence of all tags in A3 is also interesting, because it could be taken as providing yet another piece of evidence that our HAMLET data collection procedure generates increasingly more difficult data as rounds progress.

C Dataset properties

Table 9 shows the label distribution. Figure 4 shows a histogram of the number of tries per good verified example across for the three different rounds. Figure 5 shows the time taken per good verified example. Figure 6 shows a histogram of the number of tokens for contexts and hypotheses across three rounds. Figure 7 shows the proportion of different types of collected examples across three rounds.

Inter-annotator agreement Table 10 reports the inter-annotator agreement for verifiers on the dev and test sets. For reference, the Fleiss' kappa of FEVER (Thorne et al., 2018) is 0.68 and of SNLI (Bowman et al., 2015) is 0.70. Table 11 shows the percentage of agreement of verifiers with the intended author label.

D Examples

We include more examples of collected data in Table 12.

E User interface

Examples of the user interface are shown in Figures 8, 9 and 10.

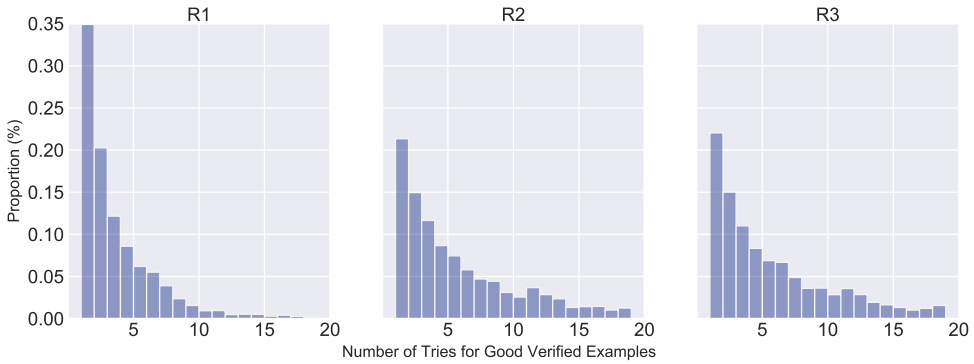

Figure 4: Histogram of the number of tries for each good verified example across three rounds.

Figure 5: Histogram of the time spent per good verified example across three rounds.

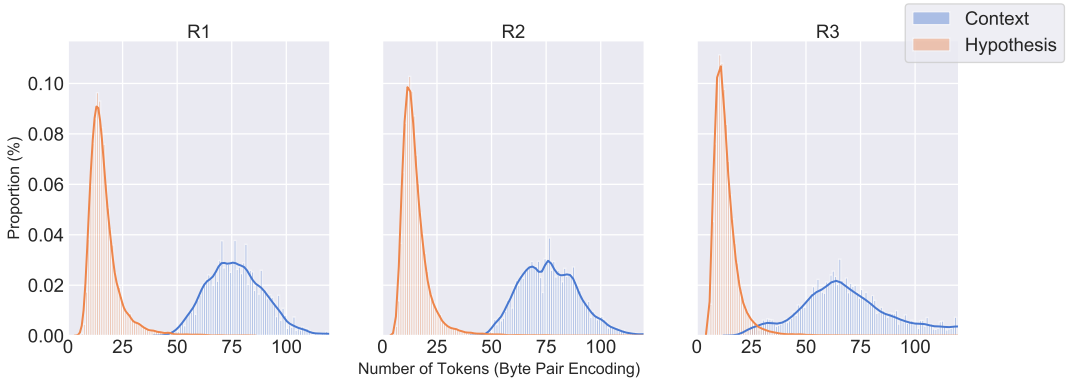

Figure 6: Histogram of the number of tokens in contexts and hypotheses across three rounds.

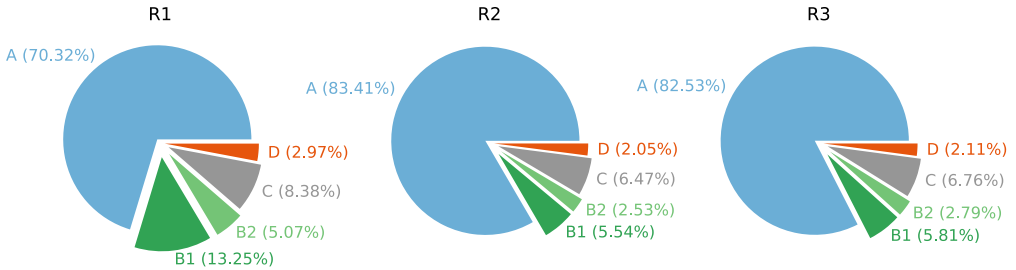

Figure 7: Proportion across three rounds. **A**=Examples that model got right, **B1**=Examples that model got wrong and the first two verifiers agreed with the writer, **B2**=Examples that model got wrong and only one of the first two verifiers agreed with the writer and a third verifier also agreed with the writer, **C**=Examples where two verifiers agreed with each other and overruled the writer, **D**=Examples for which there is no agreement among verifiers. **A** and **C** are added only to training set. **B1** and **B2** are added to training, dev, or test set. **D** was discarded.

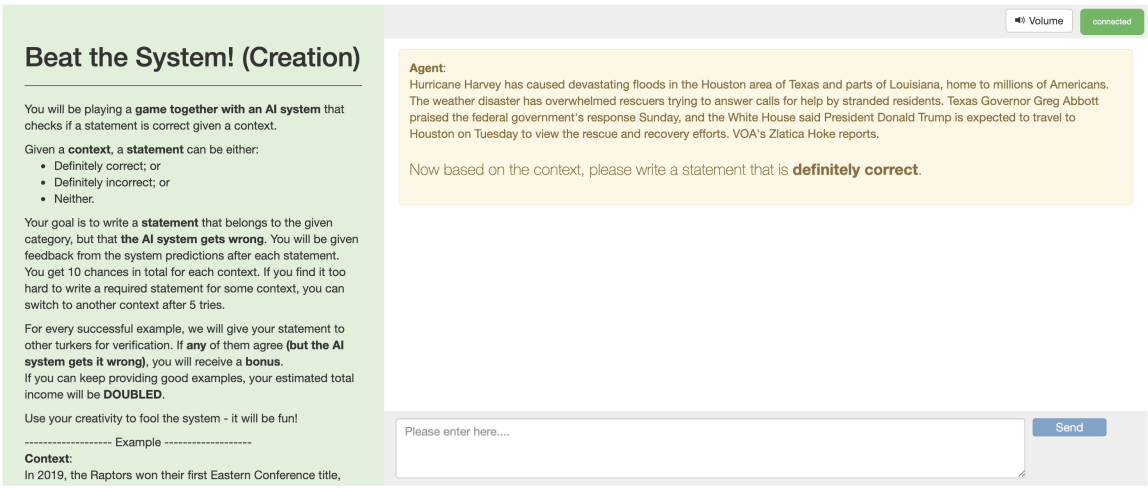

Figure 8: UI for Creation. (Provide the context to annotator)

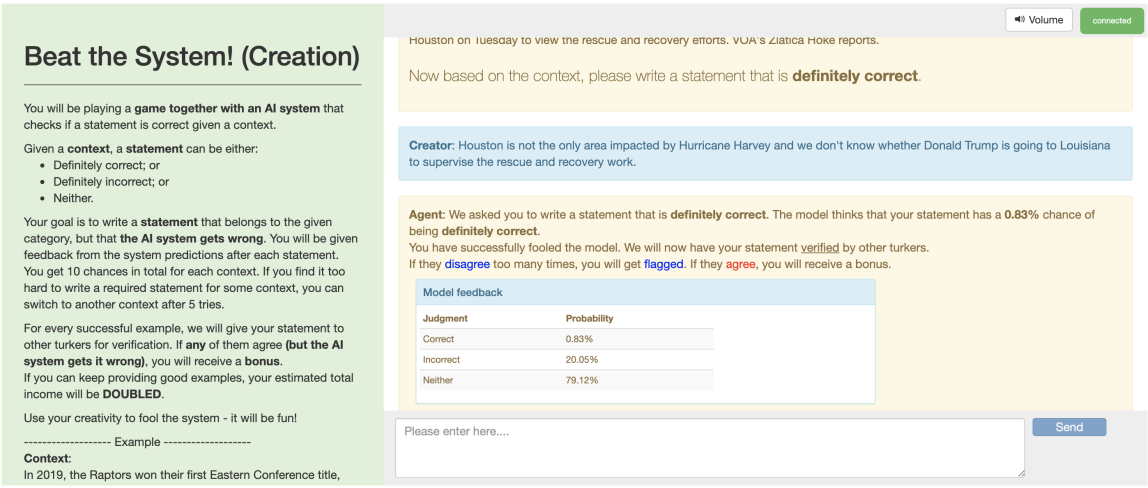

Figure 9: Collection UI for Creation. (Give the model feedback to annotator)

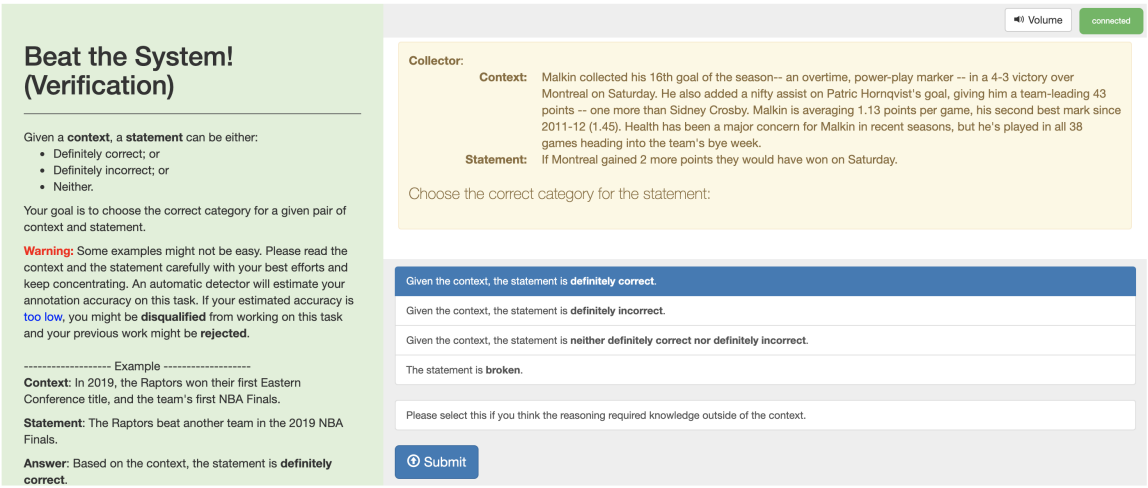

Figure 10: UI for Verification Task.

	Other Datasets							ANLI					
	SNLI		MNLI_m		MNLI_{mm}		F	A1		A2		A3	
Tag	% c	% h	% c	% h	% c	% h	% claim	% c	% h	% c	% h	% c	% h
Negation	< 1	1	14	16	12	16	3	2	6	3	10	22	*14*
'and'	30	7	41	15	42	18	6	85	*12*	88	11	75	11
'or'	1	< 1	7	2	8	2	< 1	6	0	6	< 1	15	1
Numbers	10	4	16	8	15	9	9	72	**30**	**73**	**27**	**42**	**15**
Time	12	4	15	7	16	9	6	57	**22**	**56**	**19**	**49**	**11**
WH-words	3	1	16	7	18	9	2	28	*5*	27	*5*	35	*5*
Pronouns	11	7	37	20	39	24	2	30	9	28	7	60	13
Quantifiers	5	3	21	16	22	17	3	14	10	17	*12*	38	*12*
Modals	< 1	< 1	17	13	18	14	< 1	2	3	3	2	35	*14*
$>$20 words	14	< 1	37	2	39	3	< 1	100	*5*	100	*4*	98	*4*
# exs	10k		10k		10k		9999	1k		1k		1200	

Table 8: Percentage of development set sentences with tags in several datasets: AdvNLI, SNLI, MuliNLI and FEVER. '%c' refers to percentage in contexts, and'%h' refers to percentage in hypotheses. Bolded values label linguistic phenomena that have higher incidence in adversarially created hypotheses than in hypotheses from other NLI datasets, and italicized values have roughly the same (within 5%) incidence.

	Entailment / Neutral / Contradiction		
Round	Train	Dev	Test
A1	5,371 / 7,052 / 4,523	334 / 333 / 333	334 / 333 / 333
A2	14,448 / 20,959 / 10,053	334 / 333 / 333	334 / 333 / 333
A3	32,292 / 40,778 / 27,389	402 / 402 / 396	402 / 402 / 396
ANLI	52,111 / 68,789 / 41,965	1,070 / 1,068 / 1,062	1,070 / 1,068 /1,062

Table 9: Label distribution in splits across rounds.

Round	Dev + Test	Dev	Test
A1	0.7210	0.7020	0.7400
A2	0.6910	0.7100	0.6720
A3	0.6786	0.6739	0.6832

Table 10: Inter-annotator agreement (Fleiss' kappa) for writers and the first two verifiers.

SNLI	MNLI	A1	A2	A3
85.8	85.2	86.1	84.6	83.9

Table 11: Percentage of agreement of verifiers ("validators" for SNLI and MNLI) with the author label.

Context	Hypothesis	Reason	Round	Labels orig.	pred.	valid.	Annotations
Eduard Schulte (4 January 1891 in Düsseldorf 6 January 1966 in Zürich) was a prominent German industrialist. He was one of the first to warn the Allies and tell the world of the Holocaust and systematic exterminations of Jews in Nazi Germany occupied Europe.	Eduard Schulte is the only person to warn the Allies of the atrocities of the Nazis.	The context states that he is not the only person to warn the Allies about the atrocities committed by the Nazis.	A1 (Wiki)	C	N	C C	Tricky Presupposition, Numerical Ordinal
Kota Ramakrishna Karanth (born May 1, 1894) was an Indian lawyer and politician who served as the Minister of Land Revenue for the Madras Presidency from March 1, 1946 to March 23, 1947. He was the elder brother of noted Kannada novelist K. Shivarama Karanth.	Kota Ramakrishna Karanth has a brother who was a novelist and a politician	Although Kota Ramakrishna Karanth's brother is a novelist, we do not know if the brother is also a politician	A1 (Wiki)	N	E	N E N	Standard Conjunction, Reasoning Plausibility Likely, Tricky Syntactic
The Macquarie University Hospital (abbreviated MUH) is a private teaching hospital. Macquarie University Hospital, together with the Faculty of Medicine and Health Science, Macquarie University, formerly known as ASAM, Australian School of Advanced Medicine, will integrate the three essential components of an academic health science centre: clinical care, education and research.	The Macquarie University Hospital have still not integrated the three essential components of an academic health science centre: clinical care, education and research	the statement says that the universities are getting together but have not integrated the systems yet	A1 (Wiki)	E	C	E E	Tricky Presupposition, Standard Negation
Bernardo Provenzano (31 January 1933 – 13 July 2016) was a member of the Sicilian Mafia ("Cosa Nostra") and was suspected of having been the head of the Corleonesi, a Mafia faction that originated in the town of Corleone, and de facto "capo di tutti capi" (boss of all bosses) of the entire Sicilian Mafia until his arrest in 2006.	It was never confirmed that Bernardo Provenzano was the leader of the Corleonesi.	Provenzano was only suspected as the leader of the mafia. It wasn't confirmed.	A2 (Wiki)	E	N	E E	Tricky Presupposition, Standard Negation
HMAS "Lonsdale" is a former Royal Australian Navy (RAN) training base that was located at Beach Street, Port Melbourne , Victoria, Australia. Originally named "Cerberus III", the Naval Reserve Base was commissioned as HMAS "Lonsdale" on 1 August 1940 during the Second World War.	Prior to being renamed, Lonsdale was located in Perth, Australia.	A naval base cannot be moved - based on the information in the scenario, the base has always been located in Victoria.	A2	C	N	C C	Tricky Presupposition, Reasoning Facts
Toolbox Murders is a 2004 horror film directed by Tobe Hooper, and written by Jace Anderson and Adam Gierasch. It is a remake of the 1978 film of the same name and was produced by the same people behind the original. The film centralizes on the occupants of an apartment who are stalked and murdered by a masked killer.	Toolbox Murders is both 41 years old and 15 years old.	Both films are named Toolbox Murders one was made in 1978, one in 2004. Since it is 2019 that would make the first 41 years old and the remake 15 years old.	A2 (Wiki)	E	C	E E	Reasoning Facts, Numerical Cardinal Age, Tricky Wordplay
A biker is critically ill in hospital after colliding with a lamppost in Pete The incident happened at 1.50pm yesterday in Thorpe Road. The 23-year-old was riding a Lexmoto Arrow 125 when, for an unknown reason, he left the road and collided with a lamppost. He was taken to James Cook University Hospital, in Middlesbrough, where he remains in a critical condition. Any witnesses to the collision are asked to call Durham Police on 101, quoting incident number 288 of July 9.	The Lamppost was stationary.	Lampposts don't typically move.	A3 (News)	E	N	E E	Reasoning Facts, Standard
"We had to make a decision between making payroll or paying the debt," Melton said Monday. "If we are unable to make payroll Oct. 19, we will definitely be able to make it next week Oct. 26 based on the nature of our sales taxes coming in at the end of the month. However we will have payroll the following week again on Nov. 2 and we are not sure we will be able to make that payroll because of the lack of revenue that is coming in."	The company will not be able to make payroll on October 19^{th} and will instead dispense it on October 26^{th}	It's not definitely correct nor definitely incorrect because the company said "if" they can't make it on the 19^{th} they will do it on the 26^{th}, they didn't definitely say they won't make it on the 19^{th}	A3 (News)	N	E	N C N	Reasoning Plausibility Likely, Tricky Presupposition
The Survey: Greg was answering questions. He had been asked to take a survey about his living arrangements. He gave all the information he felt comfortable sharing. Greg hoped the survey would improve things around his apartment. THe complex had really gone downhill lately.	He gave some of the information he felt comfortable sharing.	Greg gave all of the information he felt comfortable, not some. It was difficult for the system because it couldn't tell a significant difference between to word "some" and "all."	A3 (Fiction)	C	E	C C	Tricky (Scalar Implicature)

Table 12: Extra examples from development sets. 'An' refers to round number, 'orig.' is the original annotator's gold label, 'pred.' is the model prediction, 'valid.' is the validator labels, 'reason' was provided by the original annotator, 'Annotations' is the tags determined by linguist expert annotator.