WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation

Alisa Liu[♡] **Swabha Swayamdipta**[♣♢] **Noah A. Smith**[♡♣] **Yejin Choi**[♡♣]

[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Allen Institute for Artificial Intelligence [♢]University of Southern California
alisaliu@cs.washington.edu

Abstract

A recurring challenge of crowdsourcing NLP datasets at scale is that human writers often rely on repetitive patterns when crafting examples, leading to a lack of linguistic diversity. We introduce a novel approach for dataset creation based on **worker and AI collaboration**, which brings together the generative strength of language models and the evaluative strength of humans. Starting with an existing dataset, MultiNLI for natural language inference (NLI), our approach uses dataset cartography to automatically identify examples that demonstrate challenging reasoning patterns, and instructs GPT-3 to compose new examples with similar patterns. Machine generated examples are then automatically filtered, and finally revised and labeled by human crowdworkers. The resulting dataset, WANLI, consists of 107,885 NLI examples and presents unique empirical strengths over existing NLI datasets. Remarkably, training a model on WANLI improves performance on eight out-of-domain test sets we consider, including by 11% on HANS and 9% on Adversarial NLI, compared to training on the $4\times$ larger MultiNLI. Moreover, it continues to be more effective than MultiNLI augmented with other NLI datasets. Our results demonstrate the promise of leveraging natural language generation techniques and re-imagining the role of humans in the dataset creation process.
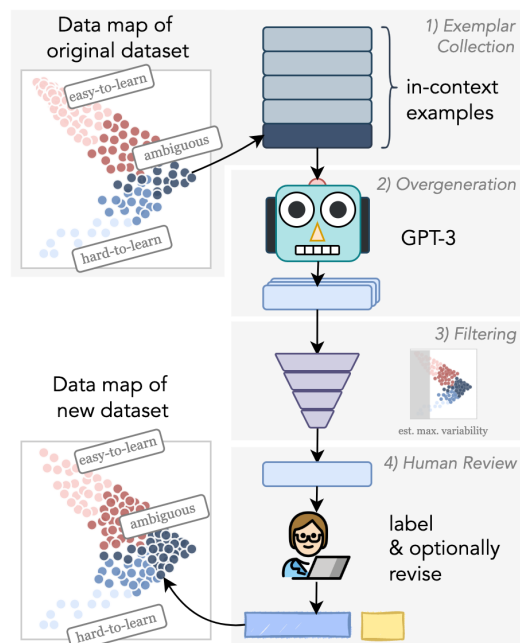
Figure 1: An illustration of our pipeline for creating WANLI. Starting with a data map (Swayamdipta et al., 2020) of an existing dataset relative to a trained model, (1) we automatically identify pockets of data instances exemplifying challenging reasoning patterns. Next, (2) we use GPT-3 to generate new instances with the same pattern. These generated examples are then (3) automatically filtered via a metric we introduce inspired by data maps, and (4) given to human annotators to assign a gold label and optionally revise.

1 Introduction

As much as large-scale crowdsourced datasets have expedited progress on various NLP problems, a growing body of research has revealed fundamental limitations in existing datasets: they are often flooded with repetitive and spurious patterns, rather than covering the broad range of linguistic phenomena required by the task (Bowman and Dahl, 2021). This leads to models that seem to achieve human-level performance on in-domain test sets, yet are brittle when given out-of-domain or adversarial examples (Ribeiro et al., 2020; Glockner et al., 2018).

We attribute this problem to an inherent challenge in the crowdsourcing design—the prevalent paradigm for creating large-scale NLP datasets—where a relatively small number of workers create a massive number of free text examples. While human annotators are generally reliable for writing *correct* examples, crafting *diverse and creative* examples at scale can be challenging. Thus, crowdworkers often resort to a limited set of writing strategies for speed, at the expense of diversity (Geva et al., 2019; Gururangan et al., 2018). When models overfit to such repetitive patterns, they fail

to generalize to out-of-domain examples where these patterns no longer hold (Geirhos et al., 2020).

On the other hand, there has been remarkable progress in open-ended text generation based on massive language models (Brown et al., 2020; Raffel et al., 2020, i.a.). Despite known deficiencies such as incoherence or repetition (Dou et al., 2021), these models often produce human-like text (Clark et al., 2021) and show potential for creative writing tasks (Lee et al., 2022). Importantly, these models are capable of replicating a pattern given just a few examples in context (Brown et al., 2020, GPT-3).

In this paper, we introduce a novel approach for dataset creation which brings together the generative strength of language models and the evaluative strength of humans through **human and machine collaboration** (§2). The key insight of our approach is that language models can create new examples by replicating linguistic patterns that are valuable for training, without necessarily "understanding" the task itself. Illustrated in Figure 1, our pipeline starts with an existing dataset. We use dataset cartography from Swayamdipta et al. (2020) to automatically identify pockets of examples that demonstrate challenging reasoning patterns relative to a trained model. Using each group as a set of in-context examples, we leverage a pretrained language model to generate new examples likely to have the same pattern (see Table 1). We then propose a novel metric, building on dataset cartography, to automatically filter generations that are most likely to aid model learning. Finally, we validate the generated examples by subjecting them to human review, where crowdworkers assign a gold label and (optionally) revise for quality.

We demonstrate the effectiveness of our approach on the task of natural language inference (NLI), which determines whether a premise entails (i.e., implies the truth of) a hypothesis, both expressed in natural language. Despite being one of the most resource-available tasks in NLP, analysis and challenge sets repeatedly demonstrate the limitations of existing datasets and the brittleness of NLI models trained on them (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018). Using MultiNLI (Williams et al., 2018) as our original dataset, we use our pipeline to create a dataset of 107,885 examples, which we call **W**orker-and-**AI NLI** (WANLI).[1]

Remarkably, empirical results demonstrate that *replacing* MultiNLI supervision with WANLI (which is 4 times smaller) improves performance on eight different out-of-domain test sets, including datasets that are converted to the NLI format from downstream tasks such as question-answering and fact verification (§3). This result holds even when augmenting MultiNLI with other NLI datasets and recently proposed augmentation sets. Moreover, including WANLI in the training data can help improve performance on certain in-domain test sets. We then analyze WANLI and show that it has fewer previously documented spurious correlations than MultiNLI (§4), and provide insights into the collaborative framework (§5).

Our approach contrasts with previous instruction-based generation of dataset examples (Schick and Schütze, 2021; West et al., 2021), which require the model to understand the task from context, fundamentally limiting the complexity of generated output to what is accessible by the model. Moreover, our human-in-the-loop approach is *collaborative*, rather than *adversarial* (Dinan et al., 2019; Nie et al., 2020; Bartolo et al., 2020). Overall, we leverage the best of both worlds: a powerful model's ability to efficiently generate diverse examples, and humans' ability to improve and ensure the quality of generations.

Our worker-AI collaborative approach is more scalable compared to the traditional crowdsourcing framework. Our approach is generalizable, allowing for rejuvenating datasets on many different classification tasks, especially when performance seems to stagnate due to overfitting to popular benchmarks (Recht et al., 2019). Our work shows the promise of leveraging language models in a controlled way to aid the dataset creation process, and we encourage the community to think of dataset curation as an AI challenge itself.

2 Worker-AI Collaborative Dataset Creation for NLI

We describe our four-stage approach for dataset creation based on worker and AI collaboration. In this work, we apply it to the task of natural language inference (NLI), which involves predicting whether a premise *entails*, *contradicts* or is *neutral* to a hypothesis. NLI has broad applicability in NLP: it has proven useful for pretraining (Clark et al., 2019; Phang et al., 2018), and can be applied to verify candidate answers in question-answering

---

[1]Pronounced wan-li like the Chinese characters 万理, as in *ten thousand reasoning*. A demo, data, and code are available at `https://wanli.allenai.org/`.

Seed MultiNLI example	Generated WANLI Example	Label & Reasoning
P: *5 percent* probability that each part will be defect free. H: Each part has a *95 percent* chance of having a defect.	P: *1 percent* of the seats were vacant. H: *99 percent* of the seats were occupied.	*Entailment* Set complements
P: The artisans, tradespeople, and providers of entertainment (reputable and not so reputable) lived downtown on the reclaimed marshlands north and east, in the area still known as Shitamachi. H: *The only place* where artisans, tradespeople and entertainers could live was in the marshlands to the north and east.	P: At the time of the Revolution, the old port of Marseille was a great center of shipbuilding and commerce. H: *The only place* where ships were built was in the old port of Marseille.	*Neutral* Hypothesis introduces exclusivity
P: To the *south*, in the Sea of Marmara, lie the woods and beaches of the Princes' Islands. H: In the *north* is the Sea of Marmara where there are mountains to climb.	P: From the park's *southern entrance*, follow the avenue *south* to the Hotel de Ville. H: From the park's *northern entrance*, follow the avenue *north* to the Hotel de Ville.	*Contradiction* Reversing cardinal directions
P: Democrats released documents indicating that Republicans sold big political donors meals with the party's leaders in federal buildings in 1995. H: *It is illegal* for a party to solicit products to donors.	P: In the late 1960s, students at a university in Wisconsin tried to organize a union. H: *It was illegal* for the students to organize a union.	*Neutral* Illegal things can happen
P: She ducked *and* parried the blow. H: She ducked *to* miss the blow.	P: She stepped on the brake *and* the car came to a stop. H: She stepped on the brake *to* stop the car.	*Entailment* Implied intention
P: To build a worldclass finance organization and help achieve better business outcomes, each of the organizations we examined *set an agenda for transforming* the finance organization by defining a shared vision -i.e. H: *The transformation was a disaster* and the entire organization had to be scrapped.	P: In order to help improve customer service, *I suggested that they send a representative* to our office to discuss our concerns. H: *The representative* sent to our office *did not solve our problems* and we lost a lot of business.	*Neutral* Intended goals may not actualize
P: Salinger *wrote* similar letters *to* other young female writers. H: Other young female writers *received* similar letters *from* Salinger as well.	P: The three schools *have* a number of students who are from families with no history of financial difficulties. H: Families with no history of financial difficulties *send* their children to the three schools.	*Entailment* Substituting a verb with a different subcategorization frame

Table 1: Seed MultiNLI examples, and corresponding WANLI examples generated by GPT-3. P stands for premise, H for hypothesis. The seed example is "ambiguous" according to the definitions of Swayamdipta et al. (2020), discussed in §2. The remaining in-context examples (shown in Appendix C.1) share the same pattern and are found using distance in `[CLS]` embeddings of a trained task model. The reasoning is a short description of the pattern we observe from the group, and which is successfully repeated in the generated example.

(Chen et al., 2021) or factuality of generated summaries (Maynez et al., 2020).

Our approach requires as prerequisites an initial dataset $\mathcal{D}_0$ and a strong task model $\mathcal{M}$ trained on $\mathcal{D}_0$. We use MultiNLI (Williams et al., 2018), a large-scale multi-genre NLI dataset, as $\mathcal{D}_0$. We finetune RoBERTa-large (Liu et al., 2019) on MultiNLI for our task model $\mathcal{M}$ (training details in Appendix B).

As an overview, we first automatically **collect** groups of examples exemplifying challenging reasoning patterns in $\mathcal{D}_0$ relative to $\mathcal{M}$, using data maps (Swayamdipta et al., 2020; Stage 1, see §2.1). Then we **overgenerate** similar examples by leveraging the pattern replication capabilities of GPT-3 (Brown et al., 2020) (Stage 2; §2.2). While GPT-3 can generate examples efficiently, it may not reliably replicate the desired pattern and its output quality will not be uniform. We address this by automatically **filtering** the generated examples using a metric derived from data maps (Stage 3; §2.3). We finally subject the collected data to **human review**, in which crowdworkers optionally revise

examples and assign gold labels (Stage 4; §2.4).

Dataset Cartography. A key component of our pipeline is inspired by data maps (Swayamdipta et al., 2020), which automatically reveal different regions in a dataset, w.r.t. the behavior of a classification model during training. These include *easy-to-learn* examples which the model consistently predicts correctly through training, *hard-to-learn* examples on which it is consistently incorrect, and *ambiguous* examples for which the model's confidence in the correct answer exhibits high *variability* across train epochs. Our pipeline focuses on *ambiguous* examples, which were shown to lead to more robust models. Additionally, ambiguous examples contain fewer spurious correlations (Gardner et al., 2021), suggesting that they capture underrepresented counterexamples to spurious correlations. Indeed, such counterexamples take more epochs of training to learn and are crucial for generalization (Tu et al., 2020), providing a potential explanation for why they appear ambiguous across early epochs and lead to more robust models.

2.1 Stage 1: Collection of Exemplars

In this stage, we automatically collect groups of examples from $\mathcal{D}_0$ which represent linguistic patterns we wish to include in the target dataset. We begin with a seed example $(x_i, y_i) \in \mathcal{D}_0$ belonging to the most ambiguous $p = 25\%$ relative to $\mathcal{M}$.[2]

To generate a new example with the same reasoning pattern, we wish to leverage the ability of GPT-3 (Brown et al., 2020) for in-context learning; hence, we need to first collect examples that test a similar kind of reasoning to x_i. To do this, we use the [CLS] token representation of each example relative to the *task* model $\mathcal{M}$, and find the $k = 4$ nearest neighbors via cosine similarity to x_i that *have the same label*. Detailed qualitative inspection shows that the nearest neighbors in this representation space tend to capture a human-interpretable similarity in the *reasoning* required to solve an example, rather than lexical or semantic similarity (examples in Table 1).

Han and Tsvetkov (2021) give another interpretation for this approach: for examples with the same label, the similarity of [CLS] token embeddings actually represents the similarity of *gradient updates* in the row of the final projection layer corresponding to that label. Thus, two examples are close if training on them would "update" the final layer of the model similarly.

By automatically identifying areas for augmentation, our method does not require any prior knowledge of challenging patterns and makes our method tractable for building on top of large-scale datasets. Nonetheless, exemplar collection could potentially be approached in different ways (e.g., through expert curation or category labels).

2.2 Stage 2: Overgeneration

Given an automatically extracted group of $k + 1$ examples from the original dataset $\mathcal{D}_0$, we construct a natural language context (prompt) for a left-to-right language model; in this work, we use GPT-3 Curie (the second-largest GPT-3 model). The prompt template we use is shown in Figure 2, where we order the examples in *increasing* similarity to the seed example.

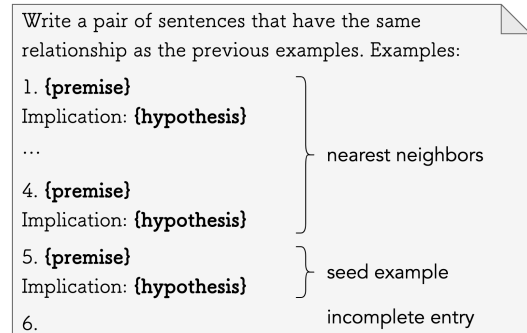Note that our method leverages GPT-3 in way

---

Figure 2: Prompt template instructing GPT-3 to generate a new example, given a set of in-context examples. To separate the premise and hypothesis, the word "Implication" is used for entailment examples (shown here), "Possibility" for neutral examples, and "Contradiction" for contradiction examples.

that is distinct from its typical usage in few-shot settings, where given examples demonstrating a task, GPT-3 performs the task on a new, unlabeled example. Here, we instead give GPT-3 examples representing a particular *slice* of the task, and ask GPT-3 to *generate* a new example in the same slice.

For each context, we sample from GPT-3 to create $n = 5$ distinct examples. We use top-p decoding (Holtzman et al., 2020) with $p = 0.5$ (additional details in Appendix C.2). Although generated examples at this stage could be assumed to share label of its $k + 1$ in-context examples, we instead consider the resulting dataset $\mathcal{D}_{\text{gen}} = \{x_i\}_i$ at the end of Stage 1 to be *unlabeled*.

2.3 Stage 3: Automatic Filtering

In this step, we wish to filter generated examples from Stage 2 to retain those that are the most ambiguous with respect to $\mathcal{M}$. However, computing ambiguity for an example requires that it be a part of the original training set, whereas we wish to estimate the ambiguity of an *unlabeled* example *without* additional training. Thus we introduce a new metric called **estimated max variability**, which measures the worst-case spread of predictions on an example x_i across checkpoints of a trained model. Let E be the total epochs in training, $\mathcal{Y}$ the label set, and $p_{\theta(e)}$ the probability assigned with parameters θ^e at the end of the e-th epoch. We define the estimated max variability as:

$$\sigma_i = \max_{y \in \mathcal{Y}} \sigma\left(\{p_{\theta(e)}(y \mid x_i)\}_{e \in E}\right), \quad (1)$$

where σ is the standard deviation function.

Concretely, we *retroactively* compute the prediction from each saved epoch of $\mathcal{M}$ on x_i. The

only assumption made is that the single example, if it had been a part of the training set, would have made a negligible difference on each model checkpoint (at least as observed through its posterior probabilities).[3] In taking a maximum across labels, we consider x_i to be ambiguous as long as $\mathcal{M}$ is undecided on *any* label $\in \mathcal{Y}$.

We first employ simple heuristics to discard examples exhibiting observable failure cases of GPT-3. Specifically, we discard examples where 1) the premise and hypothesis are identical, modulo punctuation or casing, 2) the generated example is an exact copy of an in-context example, 3) the example contains some phrases from the instruction (e.g., "pair of sentences"), or 4) the premise or hypothesis is shorter than 5 characters. Then, we compute the estimated max variability for the remaining examples with respect to $\mathcal{M}$, and retain an equal number of examples from each (intended) label class with the highest max variability, to create a dataset $\mathcal{D}_{\text{filtered}}$ that is half the size of $\mathcal{D}_{\text{gen}}$.

2.4 Stage 4: Human Review

As the final stage of our pipeline, we recruit human annotators on Amazon Mechanical Turk to review each unlabeled example $x_i \in \mathcal{D}_{\text{filtered}}$. (Details about crowdworkers and guidelines in Appendix D.) The annotator may optionally revise x_i to create a higher-quality example x_i', or let $x_i' = x_i$. Either way, they assign a label y_i. When revising examples, we asked annotators to preserve the intended meaning as much as possible through minimal revisions.[4] However, if an example would require a great deal of revision to fix *or* if it could be perceived as offensive, they should discard it. This results in the labeled dataset $\mathcal{D}_{\text{collab}} = \{(x_i', y_i)\}_i$.

Crowdworkers annotate a total of 118,724 examples, with two distinct workers reviewing each example. For examples that both annotators labeled without revision, we achieved a Cohen's κ of 0.60, indicating substantial agreement. To create the final dataset, we discard an example if *either* annotator chose to discard it, and we keep a revision only if *both* annotators revise an example (and choose a revision uniformly at random). When both annotators label the example as-is but choose different labels, we sample one of the two labels uniformly

Split	Size	Label distribution (E/N/C)
Train	102,885	38,511 / 48,977 / 15,397
Test	5,000	1,858 / 2,397 / 745

Table 2: WANLI dataset statistics.

at random. The rationale for this is discussed in Appendix D.4. This leads to a labeled dataset of 107,885 examples (90.87% of all annotated examples, with the remaining discarded). Of the labeled examples, 3.54% were revised.

We randomly split the data into a train and test sets. Key dataset statistics are summarized in Table 2. Unlike MultiNLI, WANLI is not label-balanced; see §5.3 for a discussion.

In general, we believe the role of revision depends on the quality of machine-generated examples. Indeed, we need to strike a balance between leveraging human capabilities and avoiding the re-emergence of annotation artifacts that may come with too much freedom in revision.

3 Training NLI Models with WANLI

We finetune different copies of RoBERTa-large (Liu et al., 2019) on different training sets, and evaluate each resulting model's performance on a large suite of NLI challenge sets. Given that the challenge sets were constructed independently of MultiNLI or WANLI, we consider them out-of-distribution (OOD) for both training datasets.

3.1 NLI Test Suite

The NLI challenge sets come from a wide array of domains, methodologies (e.g., crowdsourcing, expert curation, generation), and initial task formats (e.g., question-answering, fact verification).[5]

NLI Diagnostics (Wang et al., 2018) is a manually-curated test set that evaluates a variety of linguistic phenomena using naturally-occurring sentences from several domains.

HANS (McCoy et al., 2019) targets unreliable syntactic heuristics based on lexical overlap between the premise and hypothesis.

QNLI was adapted from the Stanford Question-Answering Dataset (Rajpurkar et al., 2016) by the GLUE benchmark (Wang et al., 2018). Each exam-

---

[3] Indeed, we find a high correlation between variability and estimated max variability; see Appendix A.

[4] In pilots, we found that when annotators exercised too much freedom in revision, they often re-introduced the same artifacts that have been well-documented in NLI.

[5] We evaluate on the development set for every dataset, except for Winograd NLI, where we combine the train and development set for greater statistical power, and Adversarial NLI, where we use the test set as the labels were not hidden.

	Data size	Test Set								
		Diagnostics	HANS*	QNLI*	WNLI*	NQ-NLI*	ANLI	FEVER-NLI	BIG-Bench*	WANLI
		1104	30K	5266	706	4855	3200	20K	3324	5000
MNLI	393K	68.47	78.08	52.69	56.09	62.34	32.37	68.29	64.68	64.62
MNLI + Tailor	485K	67.75	79.03	54.89	56.23	63.83	32.87	68.75	72.38	64.27
MNLI + Z-Aug	754K	66.39	80.52	57.72	55.52	62.30	33.37	68.73	66.12	64.78
MNLI ◇ ANLI	393K	67.75	79.90	68.74	60.48	62.49	54.59	72.30	72.32	**65.96**
MNLI + ANLI	556K	66.84	77.94	62.41	57.08	62.84	53.84	72.30	71.11	65.93
MNLI ◇ FEVER-NLI	393K	66.75	76.50	56.70	57.08	61.81	35.65	76.83	58.39	63.31
MNLI + FEVER-NLI	601K	67.57	76.05	52.92	54.95	63.02	35.37	76.93	64.65	64.53
MNLI + SNLI + ANLI	943K	68.75	78.65	63.38	58.49	62.94	54.21	72.02	71.05	65.10
MNLI ◇ WANLI	393K	71.01	83.10	77.00	61.89	62.94	36.46	**71.14**	76.17	75.49
MNLI + WANLI	496K	71.64	82.00	68.40	60.05	63.21	36.78	70.79	70.81	75.26
WANLI	103K	**72.73**	**89.28**	**81.40**	**67.28**	**64.18**	**41.12**	70.13	**85.19**	75.40

Table 3: Empirical comparison of different training sets for RoBERTa-large, for generalization to out-of-distribution (OOD) challenge sets. Gray cells mark settings that do not represent an OOD challenge. **Top**: Training on MultiNLI alone. **Middle**: Comparison of combination schemes with MultiNLI. We consider two data combination strategies, augmentation (+), and random replacement (◇), where the resulting dataset size is unchanged. **Bottom**: Training sets that include WANLI. The highest accuracy on each test set (excluding gray cells) is bolded. Test sets with * contain two label classes: entailment and non-entailment.

ple consists of a premise that is a sentence, and a hypothesis that is a question, which is entailed if the question is answered by the premise.

Winograd NLI was adapted by the GLUE benchmark from the Winograd Schema Challenge (Levesque et al., 2011), which tests correct coreference via common sense. To convert this dataset to NLI, an entailed hypothesis is formed by substituting a correct referent and a non-entailed hypothesis is formed by substituting an incorrect referent.

Adversarial NLI (ANLI; Nie et al., 2020) is an adversarially-constructed dataset where crowdworkers are instructed to write examples that stump existing models. Examples are collected in three rounds that progressively increase in difficulty, with model adversaries trained on MultiNLI, SNLI (Bowman et al., 2015), FEVER-NLI (discussed below), as well as ANLI sets from earlier rounds.

Natural Questions NLI (NQ-NLI, Chen et al., 2021) is created from the Natural Questions QA dataset (Kwiatkowski et al., 2019). The premise is a *decontextualized* sentence from the original context; the hypothesis consists of a question and answer candidate converted into declarative form.

FEVER NLI is adapted from the FEVER fact verification dataset (Thorne et al., 2018) and introduced along with ANLI. In each example, the premise is a short context from Wikipedia, and the hypothesis is a claim that is either supported (entailed), refuted (contradicted), or neither (neutral).

BIG-Bench NLI is a combination of four datasets from BIG-Bench (Srivastava et al., 2022) about

entailment: Analytic Entailment, Epistemic Reasoning, Disambiguation QA, Presuppositions NLI.

3.2 Training Datasets

In addition to stand-alone WANLI and MultiNLI, we also consider combining MultiNLI with other NLI datasets. We use the train sets of SNLI (Bowman et al., 2015), ANLI, and FEVER-NLI, as well as the augmentation set generated via TAILOR (Ross et al., 2022), which perturbed SNLI hypotheses to create examples with high lexical overlap between the premise and hypothesis, and the augmentation set Z-Aug (Wu et al., 2022), which was created by generating in-distribution examples and filtering them based on spurious correlations.

We consider two schemes for combining datasets $\mathcal{A}$ and $\mathcal{B}$: 1) **augmentation** ($\mathcal{A} + \mathcal{B}$), in which the two datasets are concatenated, and 2) **random replacement** ($\mathcal{A} \diamond \mathcal{B}$), where $|\mathcal{B}|$ examples from $\mathcal{A}$ are randomly swapped out and replaced with all examples from $\mathcal{B}$.

3.3 Results

Results are shown in Table 3. When comparing MultiNLI (MNLI) and WANLI alone, training a model on WANLI instead of MultiNLI leads to better performance on every test set we consider, including by 4% on Diagnostics, 11% on HANS, and 9% on Adversarial NLI. This is remarkable given WANLI is 4× smaller than MultiNLI, and contains primarily machine-written examples.

A WANLI-trained model continues to outperform baselines that combine MultiNLI with other

	Test Set				
	Diagnostics	**HANS***	**ANLI**	**BIG-Bench***	**WANLI**
ANLI	65.67	80.58	55.21	77.10	63.85
ANLI + WANLI	**72.82**	**88.58**	**56.59**	**84.89**	**75.84**

Table 4: Comparison of whether including WANLI in the training data of ANLI improves in-domain test performance, when finetuning RoBERTa-large.

NLI datasets and augmentation sets, in every OOD setting. This includes when comparing to a model trained on 9× more data from three existing NLI datasets, MNLI + SNLI + ANLI. The consistent advantage of WANLI over datasets that include ANLI (e.g., MNLI + ANLI) is noteworthy, as ANLI's adversarial creation pipeline posed a much greater challenge for human workers, and used more existing resources to train model adversaries.

Quite surprisingly, training on WANLI alone also outperforms combining WANLI with MultiNLI. This reinforces that more data might not necessarily be better, especially when the data predominantly consists of easy-to-learn examples.

In addition to the OOD setting, we consider whether augmentation with WANLI can improve *in-domain* test performance for another dataset (Table 4). Indeed, augmenting ANLI's train set with WANLI improves test accuracy on ANLI by 1.4%, while greatly aiding OOD test performance.

4 Artifacts in WANLI

We next investigate whether WANLI contains similar artifacts to MultiNLI.[6] We find that while WANLI contains fewer previously known spurious correlations, it has a distinct set of lexical correlations that may reflect artifacts in GPT-3 output.

4.1 Partial Input Models

Given that the task requires reasoning with both the premise and the hypothesis, a model that sees only one of the two inputs should have no information about the correct label. We reproduce the methodology from Gururangan et al. (2018) and train `fastText` classifiers to predict the label using partial input. After first balancing WANLI, a model trained on just the hypotheses of WANLI achieves 41.6% accuracy on the test set compared to 49.6% for MultiNLI, when restricted to the same size. A

---

[6]We note, however, that recent work has challenged whether artifacts based on partial input and lexical correlations in the dataset pose genuine robustness threats (Srikanth and Rudinger, 2022; Eisenstein, 2022).
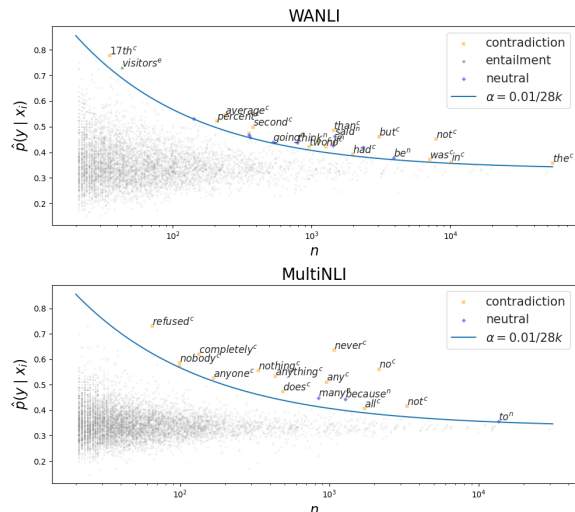

Figure 3: Competency problem-style statistical correlation plot between individual words and particular class labels, where the y-axis is the probability of label y given the presence of the word x_i, and the x-axis is the number of times word x_i appears in the data. All points representing (word, label) pairs above the blue line have detectable correlations (Gardner et al., 2021).

premise-only model trained on WANLI achieves an accuracy of 42.9%.[7]

4.2 Lexical Correlations

Gardner et al. (2021) posit that all correlations between single words and output labels are spurious. We plot the statistical correlation for every word and label in Figure 3, after balancing WANLI and downsampling MultiNLI. We observe that WANLI also contains words with detectable correlations, suggesting that GPT-3 may have some artifacts of its own due to the slightly different templates and different sets of in-context examples for each label. Interestingly, the correlations tend to be a different set of words than for MultiNLI (other than "not" and "no"), with less interpretable reasons for correlating with a certain label (e.g., "second", "was").

4.3 Premise-Hypothesis Semantic Similarity

We explore the semantic similarity between the premise and hypothesis within each label class using Sentence-BERT (Reimers and Gurevych, 2019); these distributions are shown in Figure 4. In both MultiNLI and WANLI, entailed hypotheses are naturally most semantically similar to the premise. In MultiNLI, this is followed by neutral

---

[7]Unlike WANLI, each MultiNLI premise is associated with hypotheses from all three labels; a premise-only baseline is thus guaranteed to have no information about the label.
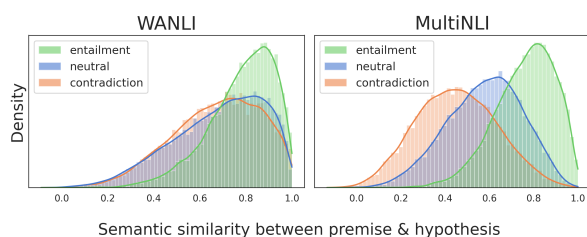
Figure 4: Semantic similarity between the premise and hypothesis, computed based on SBERT embeddings (Reimers and Gurevych, 2019). The distributions for each label class are much more well-separated in MultiNLI than in WANLI.

examples and then contradiction examples. In contrast, in WANLI there is much greater overlap in the three distributions, and those for neutral and contradiction examples are nearly indistinguishable. This suggests in WANLI, the semantic similarity between the premise and hypothesis provides less signal of the label.

5 What does WANLI show about the human machine collaboration pipeline?

We discuss observations from collecting WANLI that may shed insight for future work in the direction of collaborative dataset creation.

5.1 What kinds of revisions do annotators tend to make?

We find that revisions fall broadly into two categories: improving the fluency of the text, and improving the clarity of the relationship. The majority of revisions change the length only slightly, with 74% of both premise revisions and hypothesis revisions changing the word count between -1 and $+2$ words. Fluency revisions often target well-documented issues with text generation, such as redundancy and self-contradiction. Clarity revisions often resolve ambiguities in the example that make the entailment relationship difficult (or impossible) to determine, such as ambiguous coreference or temporal references. We provide examples of revisions in Appendix D.3.

5.2 What kinds of examples do annotators disagree on?

We find that examples on which annotators disagree provide an extremely interesting test bed for how ambiguities surface in classification tasks. Upon inspecting the examples (some are shown in Table 5), we observe that they represent genuinely ambiguous cases rather than careless mislabels, echoing

previous findings (Pavlick and Kwiatkowski, 2019). See further discussion in Appendix D.4.

5.3 How reliably does GPT-3 reproduce the in-context pattern?

One characteristic of WANLI is its imbalanced label distribution: even though the set of seed examples for generation was constructed to be balanced, after undergoing human labeling, only 15% of examples are given the contradiction label. We observe that contradiction patterns in in-context examples are generally much more challenging for GPT-3 to copy, likely because it was trained on (mostly) coherent sequences of sentences. More broadly, we find that more abstract reasoning patterns are harder for GPT-3 to mimic than patterns that involve simpler transformations.

Nonetheless, even when GPT-3 does not successfully copy the examples, the diverse set of in-context examples leads to a variety of creative output that may be challenging for human crowdworkers to achieve.

6 Related Work

Crowdsourcing The scalability and flexibility of crowdsourcing has enabled the creation of foundational NLP benchmarks across a wide range of subproblems, and made it the dominant paradigm for data collection (Mihaylov et al., 2018; Rajpurkar et al., 2016; Huang et al., 2019; Talmor et al., 2019, i.a.). Nonetheless, a growing body of research shows that resulting datasets may not isolate the key linguistic phenomena (Jia and Liang, 2017; Chen et al., 2016; Sugawara et al., 2020).

For crowdsourcing NLI datasets, where the annotator is given a premise and asked to write a hypothesis of each label (Bowman et al., 2015; Williams et al., 2018), the presence of annotation artifacts is especially well-studied (Gururangan et al., 2018; McCoy et al., 2019; Glockner et al., 2018). Recent work attempted to remedy this through different data collection protocols but found negative results (Vania et al., 2020; Bowman et al., 2020), showing this is a hard problem requiring greater innovation.

Adversarial data collection In this paradigm, annotators are asked to produce examples on which current systems fail (Kiela et al., 2021; Talmor et al., 2021; Zellers et al., 2019, i.a.). Beyond increasing annotator effort (Bartolo et al., 2020), adversarial methods have been challenged for not leading to better generalization on non-adversarial

Example	Labels	Ambiguity
P: According to the most recent statistics, the rate of violent crime in the United States has dropped by almost half since 1991. H: The rate of violent crime has not dropped by half since 1991.	*Entailment* *Contradiction*	Does "almost half" mean "not half" or "basically half"?
P: As a result of the disaster, the city was rebuilt and it is now one of the most beautiful cities in the world. H: A disaster made the city better.	*Entailment* *Neutral*	Do indirect consequences count?
P: It is a shame that the world has to suffer the pain of such unnecessary war. H: The world does not have to suffer such pain.	*Entailment* *Contradiction*	Is the scope of "has to" in the hypothesis given the war or not?
P: The original draft of the treaty included a clause that would have prohibited all weapons of mass destruction. H: The clause was removed in the final version of the treaty.	*Entailment* *Neutral*	Does the premise imply that the clause is no longer in the treaty?
P: If you can't handle the heat, get out of the kitchen. H: If you can't handle the pressure, get out of the situation.	*Entailment* *Neutral*	Is the premise to be interpreted literally or figuratively?
P: In a world of increasing uncertainty, the only certainty is that nothing is certain. H: There is no certainty in the world.	*Entailment* *Contradiction*	Self-contradictory but coherent premise

Table 5: Examples where two annotators assigned different labels. We find that many examples represent genuinely ambiguous cases rather than careless mislabels, echoing previous findings (Pavlick and Kwiatkowski, 2019).

test sets (Kaushik et al., 2021) and decreasing data diversity (Bowman and Dahl, 2021). Moreover, the resulting data has been shown to depend strongly on the adversaries, inhibiting a fair evaluation (Phang et al., 2021). Finally, these approaches may produce examples beyond the scope of the task. For example, in Adversarial NLI (Nie et al., 2020), an estimated 58% of examples required "reasoning from outside knowledge or additional facts," which is arguably separate from the underlying problem of understanding semantic entailments. We argue that we can better leverage the strengths of machines and humans by having them collaborate rather than act as adversaries.

Dataset generation Another recent approach leverages language models toward fully automatic dataset creation (Schick and Schütze, 2021; Wu et al., 2022; West et al., 2021; Bartolo et al., 2021a, i.a.). Removing human input may fundamentally limit the complexity of examples to phenomena already accessible by the model, when our goal is precisely to teach models more diverse phenomena. The most similarly-motivated work to ours, Lee et al. (2021), trains a data generator on "data-rich slices" of an existing dataset, and applies it to under-represented slices. However, they use labels or metadata to represent slices, leaving automatic methods of identifying slices to future work.

Human-machine collaboration In terms of human-machine collaboration, Tekiroğlu et al. (2020) and Yuan et al. (2021) employ a language model to generate counter-narratives to hate speech and biographies, respectively, which are validated and revised by humans. This was for a generative task, and we complement their findings by showing that human-machine collaboration can also be useful for generating labeled datasets for robust classification models. Contemporary work (Bartolo et al., 2021b) finetunes a generative annotation assistant to produce question-answer pairs that humans can revise for extractive QA.

7 Conclusion

At the heart of dataset creation is distilling human linguistic competence into data that models can learn from. The traditional crowdsourcing paradigm takes the view that the best approach for this is to solicit people to write free-form examples expressing their capabilities. In this work, we present a worker-and-AI collaborative approach and apply it to create WANLI, whose empirical utility suggests that a better way of eliciting human intelligence at scale is to ask workers to *revise* and *evaluate* content. To this end, we hope to encourage more work in developing generative algorithms to aid the dataset creation process, and therefore re-imagining the role of human annotation.

Acknowledgments

8 Ethics Statement

We acknowledge that text generated from large pre-trained language models is susceptible to perpetuating social harms and containing toxic language (Sheng et al., 2019; Gehman et al., 2020). To partially remedy this, we ask annotators to discard any examples that may be perceived as offensive. Nonetheless, it is possible that harmful examples (especially if they contain subtle biases) may have been missed by annotators and included in the final dataset. Specifically due to the above harms, we additionally caution readers and practitioners against *fully automating* any data creation pipeline.

In addition, we are cognizant of the asymmetrical relationship between requesters and workers in crowdsourcing. We took great care to pay fair wages, and were responsive to feedback and questions throughout the data collection process (see Appendix D for details). The only personal information we collect is the worker IDs from Amazon Mechanical Turk, which we will not release. The annotation effort received an IRB exemption.

9 Limitations

In this paper, we apply our collaborative dataset creation pipeline to a single language and task, English natural language inference, and leave application of the pipeline more broadly to future work.

It is possible (if not likely) that datasets partially authored by language models will have artifacts of their own, especially those reflecting social biases that may not be captured by our accuracy-based evaluation setup. For investigation of a specific generation artifact observed by Yuan et al. (2021) in their own collaborative dataset, namely the over-representation of Western entities, please see Appendix C.4.

We are not able to perform ablations on different parts of the pipeline to understand the effectiveness of each component, e.g., by comparing different means of collecting exemplar groups or different templates for prompting GPT-3. Unfortunately, such variations would be prohibitively expensive as they each require collecting a dataset of sufficient scale (along with the necessary human annotation).

Finally, although we uncover examples where annotators disagree for valid reasons (see Table 5), we only use one label per example for training and evaluation. This is because to show the effectiveness of WANLI, we need to compare WANLI to existing (singly-labeled) training datasets via performance on established (singly-labeled) benchmarks. We encourage future work to understand the limitations of forcing inherently ambiguous instances into the n-way classification scheme, or otherwise discarding these potentially valuable examples of linguistic reasoning as noise.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021a. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2021b. Models in the loop: Aiding crowdworkers with generative annotation assistants. ArXiv.

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.

T. Brown, B. Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, T. Henighan, R. Child, Aditya Ramesh, D. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, E. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

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.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text. arXiv.

Jacob Eisenstein. 2022. Informativeness and invariance: Two perspectives on spurious correlations in natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4326–4331, Seattle, United States. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Richard Zemel Claudio Michaelis, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks.

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.

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.

Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4398–4409, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwa, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *37th International Conference on Machine Learning*.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. arXiv.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, New Orleans, LA, USA.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

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*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. 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, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. 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.

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

Jason Phang, Angelica Chen, William Huang, and Samuel R. Bowman. 2021. Adversarially constructed evaluation sets are more challenging, but may not be fair. ArXiv.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. ArXiv.

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.

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):1–67.

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Neha Srikanth and Rachel Rudinger. 2022. Partial-input baselines show that NLI models can ignore context, but they don't. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4753–4763, Seattle, United States. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, and Adrià Garriga-Alonso et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *AAAI Conference on Artificial Intelligence*, pages 8918–8927.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. 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).

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2021. Synthbio: A case study in human-ai collaborative curation of text datasets. In *Neural Information Processing Systems Track on Datasets and Benchmarks*.

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.

A Estimated Max Variability

In order to test the correlation between variability and estimated max variability on a dataset $\mathcal{D}$, we would have to repeatedly hold out a single example x, train a model on $\mathcal{D} \setminus \{x\}$, and evaluate how well the estimated max variability from the model trained on $\mathcal{D} \setminus \{x\}$ correlates with the true variability from the model trained on $\mathcal{D}$, which saw x during training.

Unfortunately, this would be a very expensive experiment. Instead, we split the MNLI train set into 99% for training and 1% (3928 examples) for evaluation. For each of the held-out examples, we calculate the variability under $\mathcal{M}_{\text{MNLI}}$ and estimated max variability under $\mathcal{M}_{\text{MNLI 99\%}}$. The correlation is shown in Figure 5, and has a Pearson's correlation coefficient of 0.527 with a p-value of 7×10^{-281}.
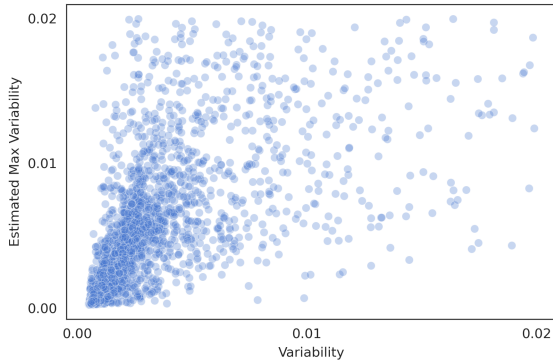
Figure 5: Correlation between variability of examples on a model that trains on the full MNLI dataset, and estimated max variability of the same examples when they are held out of the training set.

B Modeling Details

All model training is implemented with the HuggingFace (Wolf et al., 2020) library and uses the original hyperparameters from the RoBERTa paper for finetuning on GLUE (Liu et al., 2019). We train the model for five epochs and evaluate the final model. We choose not to use an early stopping scheme in order to isolate the training data as the object of study and control for training length as a confounding factor. This is important since Tu et al. (2020) showed that counter-examples can be learned better with longer training.

All training was performed on a single Nvidia Quadro RTX 6000 GPU. The duration of training varied depending on the size of the training data, from 3 hours for WANLI to 14 hours for MultiNLI + WANLI.

Hyperparameter	Assignment
Model	RoBERTa-large
Number of parameters	345M
Number of epochs	5
Learning rate	10^{-5}
Batch size	32
Weight decay	0.1
Learning rate decay	linear
Warmup ratio	0.06

Table 6: Training hyperparameters for RoBERTa-large.

C WANLI Details and Discussion

C.1 Example GPT-3 Context

We include some examples of full GPT-3 contexts in Table 12, 13, 14, 15.

C.2 GPT-3 Generation Hyperparameters

We queried the GPT-3 Curie model available through the OpenAI API[8] on the dates November 3 to November 5, 2021. In total, the generation cost $677.89. Hyperparameters for generation[9] are shown in Table 7.

Hyperparameter	Assignment
Top p	0.5
Temperature	1
Max tokens	120
Stop string	\n\n
Presence penalty	0.0
Frequency penalty	0.0

Table 7: Hyperparameters for generation from GPT-3.

C.3 Dataset sizes at each stage

In Stage 1, we collect the top 25% most ambiguous examples from each label class in MultiNLI as our set of seed examples. This leads to 98,176 seed examples, where each seed example corresponds to a unique context for GPT-3. We generate $n = 5$ examples per seed example, and skip examples that are not properly formatted with a distinct premise and hypothesis following the context template (Figure 2). At the end of Stage 2, the size of $\mathcal{D}_{\text{gen}}$ is 372,404. After applying the filtering heuristics described in §2.3 on $\mathcal{D}_{\text{gen}}$, the remaining dataset size is 287,241. Of the examples discarded, 79,278 generated examples had identical premise and hypothesis (sans punctuation and casing), and 4,732 examples had copied an in-context example. Next, we keep the half with the highest estimated max variability by sourcing an equal number of examples from each (intended) label class for a balanced dataset, resulting in $\mathcal{D}_{\text{filtered}}$ with size 143,619. However, we do not actually recruit human review on all of $\mathcal{D}_{\text{filtered}}$, and instead annotate a total of 118,724 examples. Since some of these examples are discarded, the final WANLI dataset

---

[8] https://openai.com/api
[9] described at https://beta.openai.com/docs/api-reference/completions/create

contains 107,885 examples. These correspond to 57,825 seed examples from MultiNLI.

C.4 Investigation of Western entities in WANLI versus MNLI

While we investigated known artifacts of crowd-sourced datasets in §4, generated datasets may have distinct kinds of artifacts. Indeed, recent related work qualitatively observed an over-representation of Western entities in generated biographies (Yuan et al., 2021). To investigate whether this is also characteristic of WANLI, we use flair (Akbik et al., 2019) to perform named entity recognition on MultiNLI and WANLI. Due to the challenges and ethical risks of automatically determining the origin of names and organizations, we focus on the diversity of locations mentioned. We use geopy[10] to map all locations (e.g., cities, provinces, landmarks, as well as countries) to a country.

We find that 79% of location mentions in WANLI are in Europe or North America, compared to 71% in MultiNLI. In particular, the United States is massively over-represented, accounting for 46% of mentions in WANLI and 26% in MultiNLI. However, both datasets feature a diversity of location names: WANLI mentions locations in 210 countries across 22K location entities, and MultiNLI mentions locations in 227 countries across 163K location entities. We conclude that over-representation of Western entities is indeed a concern for generated datasets, and encourage future work to consider this.

D Human Review

Screenshots of the instructions, guidelines, and annotation interface are shown in Tables 6, 7, and 8. The guidelines take inspiration from the design of the NLI Diagnostics dataset (Wang et al., 2018). To collect a pool of qualified workers, we designed a qualification task with examples testing each of these categories. NLI is a challenging task, and many generated examples are especially challenging by design. Therefore, instructing annotators in how to think about the task and resolve common issues is key to collecting high-quality, label-consistent data.

D.1 The Annotators

Annotators were required to have a HIT approval rate of 98%, a total of 10,000 approved HITs, and

---

be located in the United States.

300 Turkers took our qualification test, of which 69 passed. Turkers who were later found to produce extremely careless annotations were removed from the qualification list (and oftentimes, their annotations were discarded, though they were still paid for their work). The number of workers who contributed to the final dataset is 62.

Throughout the data collection process, the authors would review annotations and write individualized emails to Turkers with feedback, as well as group emails to clarify common challenging cases of NLI (such as examples involving questions). This follows the recommended crowdsourcing protocol from Nangia et al. (2021).

D.2 Compensation

In designing the task, we aimed for a pay rate of at least $15 per hour. Workers were paid $0.12 for each example that they annotate. At the end of data collection, we aggregate the earning and time spent from each crowdworker, and find that the median hourly rate was $22.72, with 85% of workers being paid over the $15/hour target.

D.3 Revision Analysis

We provide examples of revisions in Table 9. We find that revisions are generally targeted yet effective. The majority of revisions change the length only slightly, with 74% of both premise revisions and hypothesis revisions changing the word count between -1 and $+2$ words. A very large proportion, 11.6% of premise revisions and 20.6% of hypothesis revisions, changed the set of pronouns present in the text, often to clarify coreference.

We instructed annotators to revise examples only when it would make the example more "interesting" in some sense, or more clear without removing what's interesting. Nonetheless, we still observed a large number of revisions that greatly simplified the example, oftentimes re-introducing the same artifacts that have been documented in prior work. Therefore, we ultimately chose to include revisions only when both annotators revised the example, indicating that the revision was necessary to improve the quality of the example.

D.4 Disagreement Analysis

In order to investigate the utility of collecting a third annotation, we randomly sampled 80 examples where the two annotators disagreed on the label (and neither revised nor discarded), and two of

the authors separately annotated each one. Shockingly, the two authors agreed on the label only 49% of the time. Furthermore, in 12% of cases, all three labels were present among the four annotations. This suggests that disagreement is often due to true ambiguity rather than careless mislabeling, and a third annotation would be unlikely to have high payoff in terms of "correcting" the label. As a result, we choose not to collect a third annotation in this work. Instead, we believe that the doubly-annotated examples in WANLI have flagged many interesting cases of ambiguity in NLI, and we encourage future work to design richer annotation frameworks to uncover the source(s) of ambiguity.

We choose to keep examples with disagreement in the WANLI dataset because we believe that finetuning with one of multiple reasonable labels still provides valuable training signal.

| | | MNLI Dev. Set | |
		Matched	Mismatched
Train Set	MNLI	90.30	90.10
	MNLI ◇ WANLI	89.63	88.95
	MNLI + WANLI	89.90	89.32
	WANLI	80.17	80.46

Table 8: Results on MultiNLI's development set.

E Additional Experiments

E.1 Additional baselines

We additionally perform comparisons with several subsets of MultiNLI which are the same size as WANLI: MultiNLI filtered with the AFLite algorithm (MultiNLI with AFLite; Le Bras et al., 2020), the most ambiguous examples of MultiNLI (MultiNLI ambiguous; Swayamdipta et al., 2020), and a random subset of MultiNLI (MultiNLI downsampled). Results in Table 10 show that a WANLI-trained model outperforms these baselines on every test set.

E.2 Evaluation on MultiNLI

We report the results on MultiNLI's development set in Table 8. We find that mixing WANLI into the MultiNLI training data (either through swapping or augmentation) maintains in-domain accuracy within ∼1%. Training on WANLI alone drops performance on MultiNLI's development set by ∼10%; however, the higher performance on other out-of-domain test sets suggests that evaluation

through MultiNLI may not be a definitive signal of model ability.

E.3 Finetuning T5

We demonstrate that the robustness improvements from training on WANLI generalizes to another model architecture, T5-base (Raffel et al., 2020), which was never used in the data curation pipeline. Shown in Table 11, training T5-base on WANLI also outperforms training on MultiNLI on every test set, including by 4% of NLI Diagnostics, 10% on HANS, and 8% on Adversarial NLI (similar margins compared to finetuning RoBERTa-large).

F Data Map of WANLI

In Figure 9, we show a data map of MultiNLI relative to RoBERTa-large trained on MNLI, and of WANLI relative to RoBERTa-large trained on WANLI.

You will create high-quality examples that illustrate the relationship between two short pieces of text. Each example consists of a *premise*, a *hypothesis*, and the *relationship* between them. You will be given a *premise* and *hypothesis*, and your task is to 1) optionally revise them to improve the quality of the example, then 2) determine the relationship between them. The types of relationships are as follows.

Entailment — Given the premise, the hypothesis is definitely correct. The premise fully implies the hypothesis. For example, the premise `Pebbles the cat sat on the mat` **entails** the hypothesis `Pebbles sat` .

Contradiction — Given the premise, the hypothesis is definitely incorrect. The premise and hypothesis cannot both be true. For example, the premise `Pebbles the cat sat on the mat` **contradicts** the hypothesis `Pebbles is not on the mat` .

Neutral — Given the premise, the hypothesis may or may not be correct. The hypothesis is plausible but not entailed by the hypothesis. For example, the premise `Pebbles the cat sat on the mat` is **neutral** to the hypothesis `Pebbles purred` .

Discard — This example is low-quality or offensive in nature, and would require a great deal of revision in order to fix. In this case, there is no need to revise any text.

Before assigning a label, you may **optionally revise** the example in order to improve its quality. In these cases, you should **preserve the intended meaning** of the example as much as possible by making **minimal revisions**. Do not insert words that drastically change the meaning of the sentence, or delete entire spans of text unless they affect the fluency of the example. The goal is to ensure that the relationship is *well-defined* but *not trivially easy*; imagine you are writing **challenging** but **unambiguous** examples that could potentially be used in a classroom setting to teach or test understanding of the task.

Figure 6: Instructions provided to crowdworkers on Amazon Mechanical Turk.

Here are some guidelines to help you with determining the *relationship* between the *premise* and *hypothesis*. Remember to consult these when you are unsure.

- Presuppositions: `X knows that Y`, `X recognizes that Y`, `X shows that Y`, or `X reveals that Y` all **entail** `Y`, since `Y` is a presupposition in the premise. However, `X thinks that Y` or `X said that Y` is **neutral** with respect to `Y`, since X can be wrong. For example, `I said I would be on time` does not imply `I was on time` .
  - However, you can assume that `X said that Y` is an honest reflection of what X thinks. For example, `She said that all apples are red` **entails** `She believes that all apples are red` , and is **neutral** with respect to `All apples are red` .
- Conditionals: `If X, then Y` is **neutral** with respect to both `X` and `Y`. For example, `If the water level is low, then the engine will not start` does not imply `The water level is low` or `The engine will not start` , since the premise does not say anything about whether the water level is actually high or low!
- Background knowledge: A minimal amount of background knowledge is okay. For example, `I visited Mt. Fuji` **entails** `I visited Japan` , and `I am watching an NFL game` **contradicts** `I am watching basketball` . There may be some ambiguous cases here, and you will have to use your best judgment.
- Common sense: We should use a common sense interpretation of the text, when it strongly dominates a conflicting literal interpretation. For example, we can take `When I was young, I was obsessed with the supernatural` to **entail** `I am not obsessed with the supernatural anymore` , because it is the only commonsense way of reading the premise.
- Coreference: We can assume that expressions in the premise and hypothesis are referring to the same entity when there is a reasonable amount of corroborating information. For example, `The music building has 55 rooms` **entails** `The building has 55 rooms` and **contradicts** `The building has only one room` , by assuming "the building" in the hypothesis is referring to "the music building" in the premise. However, `The couple is talking to each other` is **neutral** with respect to `The redheads are talking to each other` , even though the couple and redheads *might* be the same two people, because there is not enough information to suggest this.
- Questions: As a rule of thumb, if the premise or hypothesis is a question (or both), consider whether saying the premise and hypothesis in sequence would add any information (**entailment**) or be contradictory (**contradiction**). For example, saying "Jane is coming at 6. When is Jane coming?" is nonsensical because the question does not need to be asked (it is already **entailed**). On the other hand, saying "Jane is coming at 6. Why isn't Jane coming?" is clearly **contradictory**. More precisely:
  - If the premise is a question and the hypothesis is a statement, we take the premise to entail its presuppositions (i.e., what is assumed in asking the question). For example, `When is Jane coming?` presupposes and therefore **entails** `Jane is coming` , and also **contradicts** `Jane is not coming` .
  - If the premise is a statement and the hypothesis is a question, it is an **entailment** if the premise answers the hypothesis, and a **contradiction** if the premise contradicts the presupposition of the hypothesis. For example, `Jane is coming at 6` **entails** `When is Jane coming?` , and **contradicts** `Why isn't Jane coming?` .
  - When the premise and hypothesis are both questions, it is an **entailment** if an answer to the premise also answers the hypothesis, and a **contradiction** if they make contradictory presuppositions. For example, `When is Jane coming?` **entails** (but is not entailed by) `Will Jane come before 6?` , and **contradicts** `Why isn't Jane coming?` (since the premise assumes Jane is coming, and the hypothesis assumes she isn't).
- Point of view: The premise and hypothesis should be read from the **same point of view**. When there is a shift in perspective that makes it seem like the premise and hypothesis are about different people, it is preferable to revise this when possible to keep the perspective consistent. For example, given the premise `I don't know if I'll ever be able to do that` and hypothesis `You can do it` , it would be preferable to revise the hypothesis to become `I can do it` . This way, the premise and hypothesis are both about `I` .

Figure 7: Guidelines provided to crowdworkers in the human review stage.

6843

1) *Premise*: He claimed that he had been pressured into giving a false confession.
Hypothesis: He had been pressured into giving a false confession.

(Optional) Revise the example below.

Premise:

He claimed that he had been pressured into giving a false confession.

Hypothesis:

He had been pressured into giving a false confession.

Given the premise, the hypothesis is...

Definitely correct	Maybe correct, maybe not	Definitely incorrect	Discard
Entailment	*Neutral*	*Contradiction*	

Figure 8: The interface on Amazon Mechanical Turk used for collecting human annotations. Annotators are given free text boxes that are pre-populated with the original premise and hypothesis, to ease the work of revision. Then, they either select an entailment class or discard the example.
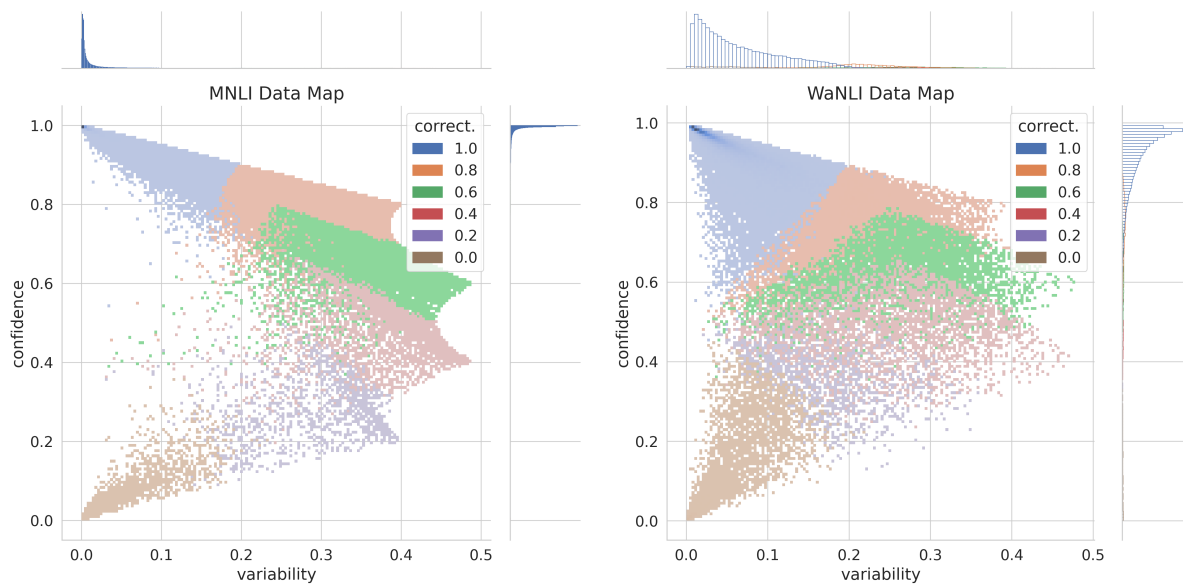

Figure 9: **Left**: Data map for MultiNLI train set, based on a RoBERTa-large classifier trained on MultiNLI. **Right**: Data map for WANLI train set, based on a RoBERTa-large classifier trained on WANLI. A comparison of the distribution in variability (which determines example ambiguity) is remarkable – we see that MultiNLI is overwhelmingly dominated by easy-to-learn examples with variability close to 0. In contrast, the distribution in variability is much more spread out in WANLI, suggesting that the dataset contains more valuable examples overall.

Example	Label	Purpose of Revision
P: ~~The power plant~~ It is the only source of continuous electric power for the city. H: The power plant is very important for the city.	*Entailment*	Coreference resolution
P: It was a well-known fact that ~~it was a well-known fact that~~ the solution was well-known. H: The solution was well-known.	*Entailment*	Redundancy
P: This will be the first time the king has met the queen in person. H: The king has met the queen in person before.	*Contradiction*	Clarity
P: She walked with a light step, as if she were floating on air. H: She was floating on air ~~, as if she were walking on air~~.	*Contradiction*	Coherence
P: There is a slight possibility that, if the same temperature data are used, the temperature of the Earth's surface in 1998 will be lower than the temperature of the Earth's surface ~~in 1998~~ now. H: The Earth's surface in 1998 was lower than the Earth's surface ~~in 1998~~ now.	*Neutral*	Self-contradiction
P: She had to go to the library to find out what the name of the street was. H: She already knew the name of the street.	*Contradiction*	Ambiguous temporal reference
P: A number of theories have been proposed to explain the decline of violence in modern society. H: Violence ~~will decline~~ has declined in modern society.	*Entailment*	Consistent tense

Table 9: Some examples of revisions that were done by annotators on examples generated by GPT-3.

		Test Set								
		Diagnostics	HANS*	QNLI*	WNLI*	NQ-NLI*	ANLI	FEVER-NLI	BIG-Bench*	WANLI
	Data size	1104	30K	5266	706	4855	3200	20K	3324	5000
MNLI	393K	68.47	78.08	52.69	56.09	62.34	32.37	68.29	64.68	64.62
MNLI (AFLite)	103K	60.50	73.73	53.91	56.37	64.28	33.12	68.04	70.75	62.19
MNLI (ambiguous)	103K	65.03	74.93	54.42	62.32	62.14	32.68	67.42	68.77	61.15
MNLI (downsampled)	103K	64.67	71.15	59.15	52.97	62.14	28.99	69.08	56.76	62.84
WANLI	103K	**72.55**	**89.40**	**76.81**	**65.15**	**64.03**	**41.12**	**70.63**	**75.40**	75.49

Table 10: Additional baselines that finetune RoBERTa-large on different subsets of MultiNLI, filtered via existing debiasing methods.

		Test Set								
		Diagnostics	HANS*	QNLI*	WNLI*	NQ-NLI*	ANLI	FEVER-NLI	BIG-Bench*	WANLI
	Data size	1104	30K	5266	706	4855	3200	20K	3324	5000
MNLI	393K	60.87	76.40	65.49	50.56	61.33	30.56	66.94	58.87	61.72
MNLI + Tailor	485K	61.14	74.34	63.33	50.70	62.05	31.06	67.15	68.95	61.28
MNLI + Z-Aug	754K	60.05	76.73	63.46	50.14	60.53	32.50	67.10	54.81	61.38
MNLI ◇ ANLI	393K	61.23	73.55	69.80	52.26	61.64	49.91	70.82	68.80	61.66
WANLI	103K	**64.58**	**86.25**	**74.66**	**51.13**	**63.66**	**38.22**	**68.27**	**76.17**	72.56

Table 11: Empirical comparison of different training datasets for T5-base. For brevity, we include MNLI, WANLI, and the strongest baselines from the results based on RoBERTa-large from Table 3.

Write a pair of sentences that have the same relationship as the previous examples. Examples:

1. In *six states*, the federal investment represents almost the entire contribution for providing civil legal services to low-income individuals.
Implication: In *44 states*, the federal investment does not represent the entire contribution for providing civil legal services for people of low income levels.

2. But if it's at all possible, plan your visit for the *spring, autumn, or even the winter*, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the *summer*.

3. *5 percent* of the routes operating at a loss.
Implication: *95 percent* of routes are operating at either profit or break-even.

4. 30 About *10 percent* of households did not
Implication: Roughly *ninety percent* of households did this thing.

5. *5 percent* probability that each part will be defect free.
Implication: Each part has a *95 percent* chance of having a defect.

6.

Table 12: Context corresponding to row 1 in Table 1, which contains *Entailment* examples from MultiNLI found via nearest neighbors in [CLS] token embedding space. All examples require reasoning about set complements, including from the universe of 100 percent, the 50 U.S. states, as well as the four seasons.

Write a pair of sentences that have the same relationship as the previous examples. Examples:

1. Small holdings abound, and traditional houses sit low on the treeless hillsides.
Possibility: The hills were *the only place* suitable to build traditional houses.

2. The inner courtyard has a lovely green and blue mosaic of Neptune with his wife Amphitrite.
Possibility: *The only colors* used in the mosaic of Neptune and Amphitrite are green and blue.

3. Nathan Road, Central, and the hotel malls are places to look.
Possibility: *The only places* to look are Nathan Road, Central and hotel malls.

4. Make your way westward to the Pont Saint-Martin for a first view of the city's most enchanting quarter, the old tannery district known as Petite France.
Possibility: *The only place* to the west of Pont Saint-Martin is the old tannery district.

5. The artisans, tradespeople, and providers of entertainment (reputable and not so reputable) lived downtown on the reclaimed marshlands north and east, in the area still known as Shitamachi.
Possibility: *The only place* where artisans, tradespeople and entertainers could live was in the marshlands to the north and east.

6.

Table 13: Context corresponding to row 2 in Table 1, which contains *Neutral* examples where the hypothesis introduces an exclusivity that is not implied by the premise.

Write a pair of sentences that have the same relationship as the previous examples. Examples:

1. Dun Laoghaire is the major port on the **_south coast_**.
Contradiction: Dun Laoghaire is the major port on the **_north coast_**.

2. Leave the city by its **_eastern Nikanor Gate_** for a five-minute walk to Hof Argaman (Purple Beach), one of Israel's finest beaches.
Contradiction: Leave the city by its **_western Nikanor Gate_** for a fifty five minute walk to Hof Argaman.

3. **_Southwest of the Invalides_** is the Ecole Militaire, where officers have trained since the middle of the 18th century.
Contradiction: **_North of the Invalides_** is the Ecole Militaire, where officers have slept since the early 16th century.

4. Across the courtyard on the **_right-hand side_** is the chateau's most distinctive feature, the splendid Francois I wing.
Contradiction: The Francois l wing can be seen across the courtyard on the **_left-hand side_**.

5. **_To the south_**, in the Sea of Marmara, lie the woods and beaches of the Princes' Islands.
Contradiction: **_In the north_** is the Sea of Marmara where there are mountains to climb.

6.

Table 14: Context corresponding to row 3 in Table 1, which contains `Contradiction` examples that flip cardinal directions between the premise and hypothesis.

Write a pair of sentences that have the same relationship as the previous examples. Examples:

1. Vendors and hair braiders are sure to **_approach_** you.
Implication: You're likely to be **_solicited by_** vendors or hair braiders.

2. The Carre d'Art, an ultramodern building opposite the Maison Carre, **_exhibits_** modern art.
Implication: Pieces of modern art **_can be found_** in the Carre d'Art, a structure which stands across from the Maison Carre.

3. But they also take pains not to dismiss the trauma the Holocaust visited and continues to visit upon Jews.
Implication: The Holocaust visited trauma upon Jews, and they are careful not to dismiss this.

4. One fortunate **_result_** of this community's influence has been the proliferation of good restaurants and interesting bars from which to choose.
Implication: The influence of this community has **_led to_** an increase in the number of intriguing bars and good dining establishments.

5. Salinger **_wrote_** similar letters **_to_** other young female writers.
Implication: Other young female writers **_received_** similar letters **_from_** Salinger as well.

6.

Table 15: Context corresponding to row 7 in Table 1, which contains `Entailment` examples that substitute a verb in the premise with one in the hypothesis that has a different subcategorization frame. Note that the third in-context example does not share quite the same pattern, but GPT-3 is still able to replicate the pattern present in other examples.