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
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 paradigm for dataset creation based on **human and machine collaboration**, which brings together the generative strength of language models and the evaluative strength of humans. Starting with an existing dataset, MultiNLI, 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 to ensure quality. The resulting dataset, WANLI, consists of 108,357 natural language inference (NLI) examples that present unique empirical strengths over existing NLI datasets. Remarkably, training a model on WANLI instead of MNLI (which is 4 times larger) improves performance on seven out-of-domain test sets we consider, including by 11% on HANS and 9% on Adversarial NLI. Moreover, combining MNLI with WANLI is more effective than combining with other augmentation sets that have been introduced. Our results demonstrate the potential of natural language generation techniques to curate NLP datasets of enhanced quality and diversity.

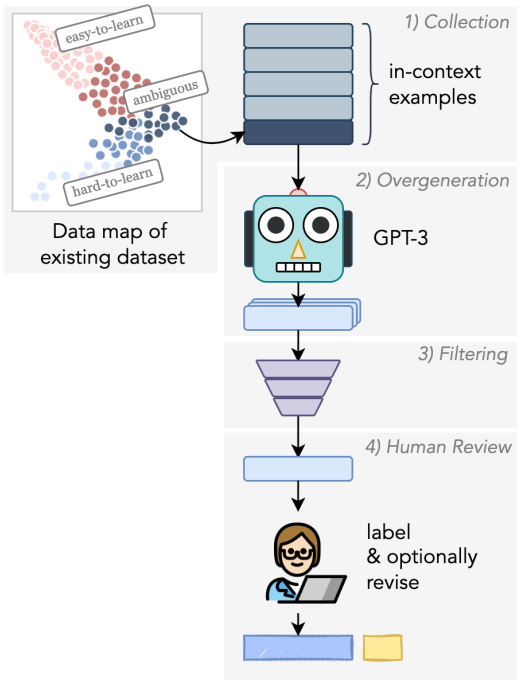

Figure 1: An illustration of our pipeline for creating WANLI. Starting with a data map (Swayamdipta et al., 2020) of an existing dataset relative to a trained model, (1) we automatically identify pockets of data instances exemplifying challenging reasoning patterns. Next, (2) we use GPT-3 to generate 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 or 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; Jia and Liang, 2017; 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, crowd-

workers 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 are increasingly being deployed in creative writing tasks (Alex, 2020). Importantly, these models also demonstrate a remarkable capability to replicate a pattern given just a few examples in context (Brown et al., 2020, GPT-3).

In this paper, we introduce a novel paradigm 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, then automatically identifies pockets of examples (see Table 1) that demonstrate challenging reasoning patterns relative to a trained model, using dataset cartography from Swayamdipta et al. (2020). 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. We propose a novel metric, also inspired by data maps, to automatically filter generations for those 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 108,357 examples, which we call **W**orker-and-**AI** **NLI** (WANLI; §3).[1]

Remarkably, empirical results demonstrate that *replacing* MultiNLI supervision with WANLI (which is 4 times smaller) improves performance on seven 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 (§4). Moreover, under a data augmentation setting, combining MultiNLI with WANLI is more effective than using other augmentation sets that have been introduced. Finally, including WANLI in the training data can help improve performance on certain in-domain test sets. We also analyze spurious correlations in WANLI and find that it has fewer previously documented artifacts than MultiNLI (§5). Finally, we provide insights into the collaborative framework based on our NLI data collection (§6).

Our approach contrasts with previous instruction-based generation of dataset examples (Schick and Schütze, 2021; West et al., 2021), which requires 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*, since we do not ask annotators to write examples specifically to challenge a model (Nie et al., 2020; Le Bras 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 discriminate the quality of generations.

Overall, our worker-AI collaborative pipeline is scalable compared to a crowdsourcing framework requiring human authorship. The generality of our approach allows it to be applied 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.

We will release the data and code soon.[2]

[1] Pronounced wan-li like the Chinese characters 万理, as in *ten thousand reasoning*.

[2] Please email the corresponding author for interim data requests.

Seed MNLI 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 of action
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 MNLI examples, and corresponding WANLI examples which were fully 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) 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.

2 Worker-AI Collaborative Dataset Creation

We outline a four-stage pipeline for dataset creation, which combines the capabilities of powerful language models and human annotators. Our prerequisites include an initial dataset $\mathcal{D}_0$ and a state-of-the-art model $\mathcal{M}$ trained on $\mathcal{D}_0$.

Our goal is to create examples which would enhance model robustness on unseen distributions. We facilitate this by automatically **collecting** 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 automatically **overgenerate** similar examples, 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 (e.g., the original label may not be correct) and the quality of its output 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 assign gold labels and optionally revise examples for quality (Stage 4; §2.4).

Dataset Cartography A key component of our pipeline relies on dataset cartography (Swayamdipta et al., 2020), a method that discovers different regions in a dataset w.r.t. the behavior of a model during training. Data maps reveal model-specific regions in the data, including *easy-to-learn* examples which the model consistently and correctly predicts throughout training, and *hard-to-learn* examples on which it is consistently incorrect. Specifically, we are interested in examples from the third, *ambiguous* region of the map, for which the model's confidence exhibits high *variability* in the correct answer across epochs of training. Gardner et al. (2021) showed that ambiguous examples in a dataset contain fewer spurious correlations, suggesting that ambiguity captures under-represented 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 result in more robust models. Finally, ambiguous examples have demonstrated desirable properties from an information-theoretic perspective (Ethayarajh et al., 2021).

2.1 Stage 1: Collection of Exemplars

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 $p\%$ relative to $\mathcal{M}$, within each label class for a balanced seed dataset. We use the `[CLS]` token representation of each example relative to the *task* model $\mathcal{M}$, and find the k-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 that are candidates 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, we note that exemplar collection could potentially be approached in different ways (e.g., through expert curation, or human-written rationales).

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. Note that our method leverages GPT-3 in way that is distinct from its traditional 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 within the same slice. While the prompt consists of examples that share a label, the label is not necessarily shown to the language model. For each context, we sample from GPT-3 to create n distinct examples. 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 the task model $\mathcal{M}$. However, computing ambiguity for a generated example requires that the example be a part of the original training set, and have a gold label. Instead, we wish to estimate the ambiguity of our unlabeled examples without any additional training. We introduce a new metric called **estimated max variability**, which measures the worst-case spread of predictions on an example x_i across epoch-specific 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. Then, we define the estimated max variability as:

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

where σ indicates standard deviation.

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). By taking a maximum across labels, the intuition is that as long as $\mathcal{M}$ is undecided on any particular label, we consider the example to be ambiguous. We perform a case study on MNLI and show that there is high correlation between the original variability metric and estimated max variability in Appendix A.

Our resulting dataset, $\mathcal{D}_{\text{filtered}} \subseteq \mathcal{D}_{\text{gen}}$ retains the top $q\%$ of $\mathcal{D}_{\text{gen}}$ within each (intended) label class,[3] based on our max variability metric.

2.4 Stage 4: Human Review

As the final stage of our pipeline, we recruit human annotators to review each unlabeled example $x_i \in$

[3] $\mathcal{D}_{\text{filtered}}$ is still unlabeled in Stage 3.

$\mathcal{D}_{\text{filtered}}$. The annotator may optionally revise x_i to create a higher-quality example x'_i, or let $x'_i = x_i$. Either way, the annotator assigns a label y_i. This results in our final dataset $\mathcal{D}_{\text{collab}} = \{(x'_i, y_i)\}_i$.

Broadly, 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 WANLI

We implement our proposed paradigm (§2) by creating a new dataset for natural language inference (NLI). The task involves predicting whether a premise statement entails, contradicts or is neutral to a hypothesis statement. Despite great progress on the task, NLI models still struggle on out-of-domain instances, suggesting they have overfit to existing datasets (Gururangan et al., 2018; Poliak et al., 2018).

As our initial dataset $\mathcal{D}_0$, we use MultiNLI (Williams et al., 2018), a large-scale multi-genre dataset for the NLI task. [4] We finetune RoBERTa-large (Liu et al., 2019) on MultiNLI for our task model $\mathcal{M}$.

3.1 Stage 1: Collection of MultiNLI Exemplars

As seed examples, we use MultiNLI examples that are in the top $p = 25\%$ for ambiguity w.r.t RoBERTa-large. For each seed example, we collect the $k = 4$ nearest same-label neighbors based on the `[CLS]` embedding from $\mathcal{M}$, to create a total of 5 in-context examples for GPT-3. We order the examples in *increasing* similarity to the original example, so that it is the last example in the context. We use a slightly different instructional template for each label class, which is shown in Figure 2.

3.2 Stage 2: Overgeneration

We use GPT-3 Curie (the second-largest model from the OpenAI API[5]) and use top-p decoding (Holtzman et al., 2020), where $p = 0.5$. For each set of in-context examples, we sample $n = 5$ new examples.

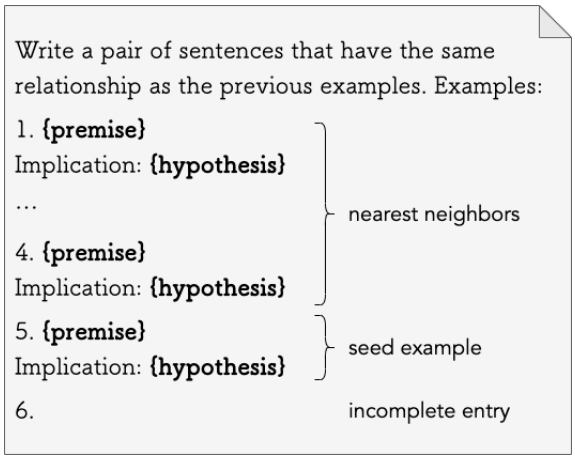

Figure 2: Template for prompting GPT-3 to generate a new example, given a set of in-context examples. To separate the premise and hypothesis, the word "Implication" is used for entailment examples, "Possibility" for neutral examples, and "Contradiction" for contradiction examples. Actual examples of complete prompts are shown in Appendix C.

3.3 Stage 3: Automatic Filtering

Before applying our estimated max variability metric to filter GPT-3 generations (§2.3), we employ simple heuristics to discard examples exhibiting observable failure cases of GPT-3. Concretely, we discard examples where 1) the premise or hypothesis is shorter than 5 characters, 2) the premise and hypothesis are identical, sans punctuation or casing, 3) the generated example is an exact copy of an in-context example, 4) the example contains some phrases from the instruction (e.g., "pair of sentences"). Then, we compute the estimated max variability for the remaining examples, with respect to M. We retain generations from the top $q = 50\%$ of estimated max variability.

3.4 Stage 4: Human Review

We use the Amazon Mechanical Turk platform for human annotation. We have a total of 62 crowdworkers, who are given a thorough set of NLI guidelines, and undergo qualification tests and further training (details in Appendix D).

Given an unlabeled example, annotators are asked to optionally revise it to be higher-quality, while preserving the intended meaning as much as possible through minimal revisions. In initial pilots, we found that when annotators exercised too much freedom in revision, they often re-introduced

[4] We exclude the telephone genre of MultiNLI, which consists of transcripts of telephone conversations, due to their extremely low fluency and ill-defined entailment relationships. During pilots, we found that generated examples mimicking telephone conversations would require human crowdworkers to read low-quality text and revise for basic fluency, without providing high-quality NLI examples.

[5] https://openai.com/api/

Split	Size	Label distribution (E/N/C)
Train	103,357	39,103 / 48,816 / 15,438
Test	5,000	1,853 / 2,386 / 761

Table 2: WANLI dataset statistics.

the same artifacts that have been well-documented in NLI. After revising, annotators assign a label among the three entailment classes. Alternatively, if an example would require a great deal of revision to fix, they may simply discard it.

For data quality, two distinct annotators label each example in $\mathcal{D}_{\text{filtered}}$. For examples that both annotators labeled without revision, we achieved a Cohen Kappa score of 0.60, indicating substantial agreement. Disagreements are discussed in §6.2, with examples.

To create the final dataset $\mathcal{D}_{\text{collab}}$, we discard an example if *either* annotator chose to discard it, and keep a revision only if *both* annotators believed that it needed revision (and we 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 at random.

Crowdworkers annotate a total of 118,724 examples. Of these, 108,357 (91.27%) are labeled and presented in $\mathcal{D}_{\text{collab}}$ and the remaining 8.73% are discarded. Of the labeled examples, 3.43% were revised. We randomly split the data into a train set and test set. Key dataset statistics are summarized in Table 2.

4 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 the challenge sets were constructed independently of MultiNLI or WANLI, we consider them out-of-distribution for both training datasets.

4.1 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). 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 (Nie et al., 2020), where we use the test set because the labels are not hidden.

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 example consists of a premise that is a sentence, and a hypothesis that is a question. The example is entailed if the premise contains an answer to the question, and non-entailed otherwise.

Winograd NLI was adapted from the Winograd Schema Challenge (Levesque et al., 2011) by the GLUE benchmark. In each example, the premise is a sentence containing a pronoun, whose referent is not explicitly stated but easily inferable with common sense (e.g., "The trophy did not fit in the suitcase because it is too big."). To convert this dataset to NLI, an entailed hypothesis is formed by substituting the correct referent for the pronoun; a non-entailed hypothesis is formed by substituting the incorrect referent.

Adversarial NLI (ANLI; Nie et al., 2020) is an adversarially-constructed dataset where crowdworkers are instructed to write examples that existing models cannot answer correctly. Examples are collected in three rounds that progressively increase in difficulty and complexity, 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 (rewritten in order to stand alone), and the hypothesis consists of a question and answer candidate which is converted into declarative form.

FEVER NLI is adapted from the FEVER fact verification dataset (Thorne et al., 2018), and introduced along with the Adversarial NLI benchmark. 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).

		Test Set									
		Diagnostics	HANS*	QNLI*	WNLI*	NQ-NLI*	Adversarial NLI			FEVER-NLI	WANLI
							R1	R2	R3		
	Dataset size →	1104	30K	5266	706	4855	1000	1000	1200	20K	5,000
Training Set	MultiNLI	68.47	78.08	52.69	56.09	62.34	47.49	26.10	25.00	68.29	64.62
	WANLI	**72.73**	**89.80**	**78.19**	**66.57**	**64.40**	**48.80**	**36.39**	**39.83**	**69.52**	75.48
	MultiNLI + Tailor	67.75	79.03	54.89	56.23	63.83	46.99	27.70	25.41	68.75	64.27
	MultiNLI ⋄ ANLI	67.75	79.90	68.74	60.48	62.49	72.69	47.20	45.66	72.30	65.96
	MultiNLI + ANLI	66.84	77.94	62.41	57.08	62.84	71.20	47.20	44.91	72.30	65.93
	MultiNLI ⋄ FEVER-NLI	66.75	76.50	56.70	57.08	61.81	54.19	28.49	26.16	76.83	63.31
	MultiNLI + FEVER-NLI	67.57	76.05	52.90	54.95	63.02	**54.69**	28.49	25.00	76.93	64.53
	MultiNLI ⋄ WANLI	**72.01**	**83.80**	**71.53**	**61.61**	**64.07**	51.80	**30.70**	**29.33**	**71.20**	75.05
	MultiNLI + WANLI	71.10	82.90	69.52	59.77	63.66	50.00	29.39	27.83	70.91	74.69
	ANLI	65.67	80.58	**81.25**	66.00	62.03	72.79	47.79	46.75	72.74	63.85
	WANLI + ANLI	**71.92**	**88.81**	80.93	**67.13**	**63.83**	**73.60**	**50.19**	**48.00**	**73.59**	**75.84**

Table 3: Experimental results comparing different training sets for RoBERTa-large. The rows represent training sets, and the columns represent test sets. Test sets with * indicate that they contain two label classes: entailment and non-entailment. We consider two data combination strategies, 1) augmentation (denoted by +), and 2) random replacement (denoted by ⋄), in which the resulting dataset size does not change. **Top**: Comparison of MultiNLI and WANLI as the standalone training sets. **Middle**: Comparison of combination schemes with MultiNLI. In the top two sections, we compare generalization to out-of-domain challenge sets; gray cells mark settings that do not represent an out-of-domain challenge. **Bottom**: Comparison of whether including WANLI in the training data improves performance on in-domain test data. Within each section, the highest accuracy on each test set (excluding gray cells) is bolded.

4.2 Training Datasets

In addition to stand-alone WANLI and MultiNLI, 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 examples from $\mathcal{B}$, such that the overall size of the dataset is unchanged from $\mathcal{A}$ alone. Under the combination schemes, we compare to other NLI datasets introduced to target the limitations of existing NLI datasets. In particular, we use the train sets of ANLI and FEVER as well as the augmentation set generated via TAILOR (Ross et al., 2021), which used linguistic perturbation strategies on SNLI hypotheses (Bowman et al., 2015) to create examples with high lexical overlap between the premise and hypothesis.

Finally, we investigate whether combining WANLI with ANLI can help improve in-domain performance on ANLI.

4.3 Results

Results are shown in Table 3. When comparing MultiNLI and WANLI alone, a RoBERTa-large model trained on WANLI performs better than a RoBERTa-large MultiNLI model on every test set we evaluate on, including by 4% on Diagnostics, 11% on HANS, and 9% on Adversarial NLI (over three rounds). This is remarkable considering the smaller size of WANLI (by a factor of 4) and the fact that examples are dominantly machine-written.

Perhaps surprisingly, training on WANLI alone performs consistently better than combining WANLI with MultiNLI; moreover, of the two combination settings, replacement with WANLI (MultiNLI ⋄ WANLI) is better than augmentation (MultiNLI + WANLI). This reinforces that more data is not better, especially when it comprises predominantly of easy-to-learn examples. Nonetheless, both MNLI + WANLI and MNLI ⋄ WANLI improve performance upon the baseline MultiNLI-trained model for every test set. In addition, WANLI is more valuable than Adversarial NLI on every test set other than ANLI's own test set and FEVER-NLI, which was used to train adversaries for the ANLI dataset creation process. This result is substantial because the creation pipeline of Adversarial NLI, which required annotators to craft examples that fool existing models, posed a much greater challenge for human workers and used more existing resources to train adversaries. For MultiNLI + Tailor, we see a modest improvement on HANS ($\sim 1\%$), which also targets heuristics related to high lexical overlap.

We then consider whether WANLI can further improve performance on ANLI by using the corresponding training set. Indeed, augmenting ANLI's

train set with WANLI improves test accuracy upon ANLI by 1.5%, and upon MultiNLI ⋄ ANLI (the next best setting) by 2%. This shows that WANLI may also be used to improve performance on in-domain test sets, while also greatly improving out-of-domain test performance.

5 Artifacts in WANLI

We next seek to answer whether WANLI contains similar annotation artifacts to MultiNLI. We find that while WANLI contains fewer previously known spurious correlations, it contains a different set of lexical correlations that may reflect artifacts in GPT-3 output.

5.1 Partial Input Models

Given that the nature of the task involves 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. To test this, we reproduce the methodology from Gururangan et al. (2018) and train two `fastText` classifiers to predict the label using only the premise or only the hypothesis as input. After first balancing WANLI, a model trained on just the hypotheses of WANLI achieves 41.6% accuracy compared to 49.6% for MultiNLI, when restricted to the same size. A premise-only model trained on WANLI achieves an accuracy of 42.9%.[6]

5.2 Lexical Correlations

Gardner et al. (2021) propose the competency problems framework, which posits that all correlations between single words and output labels are spurious. In other words, no single word should give information about the class label. Following Gardner et al., 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 many words with detectable correlations, suggesting that GPT-3 may have some artifacts of its own due to the slightly different templates for each label, and the 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"), with potentially less interpretable reasons for correlating with a certain label (e.g., "whether", "was").

[6]Because each premise in MultiNLI is associated with a hypothesis from each entailment class, a premise-only baseline is guaranteed to have no information about the output label.

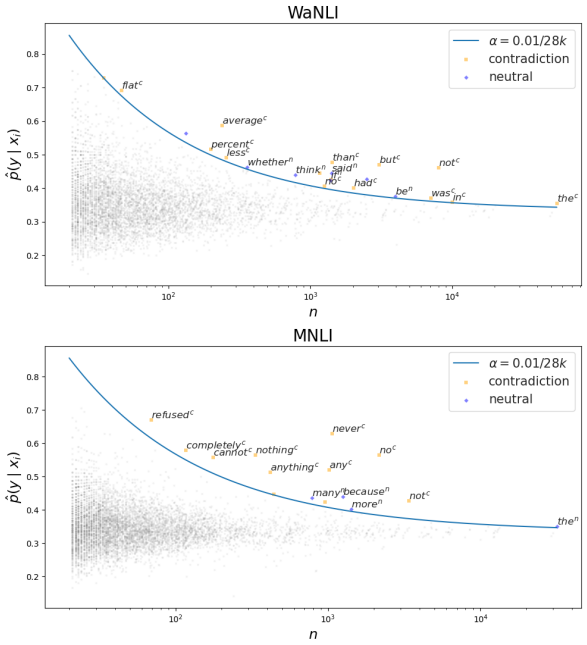

Figure 3: Competency problem-style statistical correlation plot between individual words and particular class labels, where the y-axis is the probability of label y given the presence of the word x_i, and the x-axis is the number of times word x_i appears in the data. All points representing (word, label)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

5.3 Semantic Similarity between Premise & Hypothesis

We additionally explore the semantic similarity between the premise and hypothesis within each entailment class. We separately encode the premise and hypothesis with Sentence-BERT (Reimers and Gurevych, 2019), which computes semantically meaningful text embeddings that can be compared using cosine similarity. The distribution of premise-hypothesis similarity within each entailment class is 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 examples and then contradiction examples. In contrast, in WANLI, there is much greater overlap in the three distributions, and the distributions 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, and may explain the difficulty of the benchmark.

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: The Commission did not consider the costs of this rule. H: The rule will not cost anything.	*Contradiction* *Neutral*	Does "considering the costs" imply that the costs are non-zero?
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: He'd made it clear that he was not going to play the game. H: He didn't want to play the game.	*Contradiction* *Neutral*	Can we assume intention behind actions?
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: After two hours of discussion, the group decided to meet again the next day. H: The group will meet again on the next day.	*Entailment* *Neutral*	Can we assume follow-through on a decision?
P: He felt as if he were watching a movie and was having a hard time distinguishing between the actors and the real people. H: He was watching a movie and could not tell the difference between the actors and the real people.	*Entailment* *Contradiction*	Is the hypothesis a metaphorical statement?
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?

Table 4: 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).

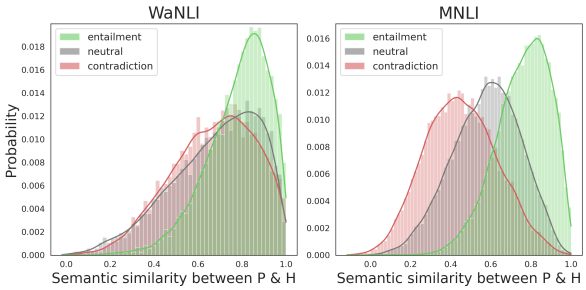

Figure 4: Semantic similarity between the premise and hypothesis, computed based on Sentence-BERT embeddings (Reimers and Gurevych, 2019). The distributions for each entailment class are much more well-separated in MultiNLI than in WANLI.

6 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.

6.1 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 entailment 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.

6.2 What kinds of examples do annotators disagree on?

We find that examples on which annotators disagree provide an extremely interesting test bed for analyzing inherent disagreements in NLI (see Table 4). Upon inspecting the examples on which annotators disagreed, we observe that many examples represent genuinely ambiguous cases rather than careless mislabels, echoing previous findings (Pavlick and Kwiatkowski, 2019). Future work could further investigate these disagreements, and learning strategies that take into account the inher-

ent ambiguity of these examples.

6.3 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 15.35% 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.

7 Related Work

Adversarial filtering There has been a trend toward adversarial data collection paradigms, in which annotators are asked to produce examples on which current systems fail (Le Bras et al., 2020; Kiela et al., 2021; Talmor et al., 2021; Jia and Liang, 2017; Zellers et al., 2019, i.a.). While this eliminates examples containing artifacts that models have already learned, it does not prevent the creation of new ones on which current models fail. More importantly, these datasets may obscure the abilities we actually want to measure by systematically eliminating coverage of linguistic phenomena that are already well-solved by the adversary (Bowman and Dahl, 2021). Finally, these approaches may result in 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 appears different from the underlying NLI task of understanding semantic entailments. We believe that we can better leverage the strengths of machines and humans by having them collaborate rather than act as adversaries.

Dataset generation Recently, there have been approaches that leverage the generative ability of pretrained language models toward fully automatic dataset generation (Schick and Schütze, 2021; Anonymous, 2021; Lee et al., 2021; West et al., 2021). In contrast, our work is unique in its method for automatically identifying groups of valuable examples, as well as using human annotation as the final stage in the pipeline. The most similarly-motivated work, Lee et al. (2021), trains a generator to simulate data creation on "data-rich slices" of the data, and applies it to under-represented slices. However, they use labels or meta-data to represent particular slices of task, leaving automatic methods of identifying slices to future work. Also in contrast, we leverage an off-the-shelf language model rather than finetuning a generator on existing data, with the intuition that this would induce greater diversity in generations.

In terms of human-machine collaboration, Tekiroğlu et al. (2020) employ a language model to generate counter-narratives to hate speech, which are validated and revised by experts. 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.

8 Conclusion

We present a general paradigm for crowdworker and AI collaboration for creating large-scale datasets, as well as a particular instantiation of it for natural language inference: WANLI. We show that WANLI improves generalization to out-of-distribution settings while avoiding known issues in existing NLI datasets, such as MultiNLI.

More broadly, while the crowdsourcing paradigm takes the view that the best way to distill human capabilities is by soliciting people to write free-form examples expressing their linguistic capabilities, our work suggests that a better way of eliciting human intelligence at scale is by asking workers to *revise* and *evaluate* content. To this end, we hope to encourage more work in developing methods of leveraging advances in large pretrained language models to aid the dataset creation process. Future directions involve exploring adaptations of our collaborative framework for collecting datasets in truly low-resource settings.

References

Vlad Alex. 2020. 20 creative things to try out with gpt-3. Accessed: 2022-01-14.

Anonymous. 2021. Generating data to mitigate spurious correlations in natural language inference datasets. Open Review.

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.

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)*.

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.

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.

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

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2021. Information-theoretic measures of dataset difficulty.

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.

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*.

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.

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.

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*.

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.

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.

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 E. Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. arXiv.

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.

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, 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.

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.

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 $\mathcal{D} - \{x\}$, and evaluate how well the estimated max variability from the model trained on $\mathcal{D} - \{x\}$ 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% for evaluation, corresponding to 3928 examples. 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}.

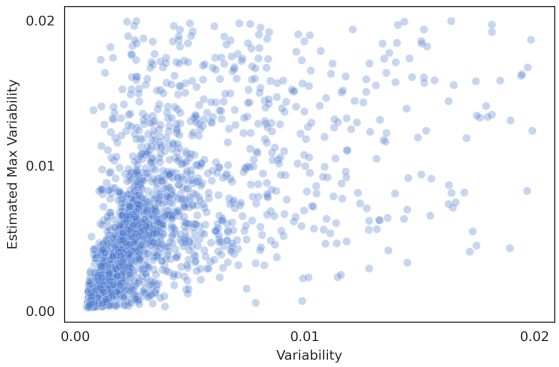

Figure 5: Correlation between variability of examples on a model that trains on the full MNLI dataset, and estimated max variability of the same examples when they are held out of the training set.

B Modeling Details

We use the original hyperparameters from the RoBERTa paper for finetuning on GLUE. For each dataset, 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.

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 5: Training hyperparameters for RoBERTa-large.

C GPT-3 Context Examples

We include some examples of full GPT-3 contexts in Table 7, 8, 9, 10.

D Human Annotation

We provide guidelines as shown in Figure 6, taking inspiration from the design of the NLI Diagnostics dataset (Wang et al., 2018). In addition, we designed a qualification task with examples testing understanding of each of these categories. We believe that 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.

A screenshot of the revision and labeling interface is shown in Figure 7.

D.1 Revision Examples

We provide examples of revisions in Table 6. We observed that a very large proportion, 11.6% of premise revisions and 20.6% of hypothesis revisions, changed the set of pronouns present in the text. While we worked very hard to instruct annotators to revise examples only when it would make the example more "interesting" in some sense, or more clear without removing what's interesting, unfortunately 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 (Gururangan et al., 2018). Therefore, we ultimately chose to include revisions only when both annotators revised the example.

D.2 Disagreement Examples

We provide examples where the two annotators disagreed in Table 4. These examples demonstrate that language models are capable of generating extremely challenging examples. Although we originally considered recruiting a third annotation where the two labels disagreed, upon reviewing many cases of disagreement, we found that a large proportion were truly ambiguous rather than careless mislabels, or the reasoning required was so subtle that a third annotator may not resolve the tie correctly. Therefore, a third annotation would likely have low reward in terms of "fixing" annotations.

E Data Map of WANLI

In Figure 8, we show a data map of MultiNLI relative to RoBERTa-large trained on MNLI, and of WANLI relative to RoBERTa-large trained on WANLI.

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: I've never been able to figure out how the system works. H: I still don't know ~~The system is~~ how the system works.	*Entailment*	Coherence
P: This year's spring break was a disaster for most of the students. H: The students were not all able to have a good time during spring break.	*Entailment*	Clarity
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 6: Some examples of revisions that were done by annotators on examples generated by GPT-3.

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 6: Guidelines provided to crowdworkers in the human review stage.

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 7: 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.

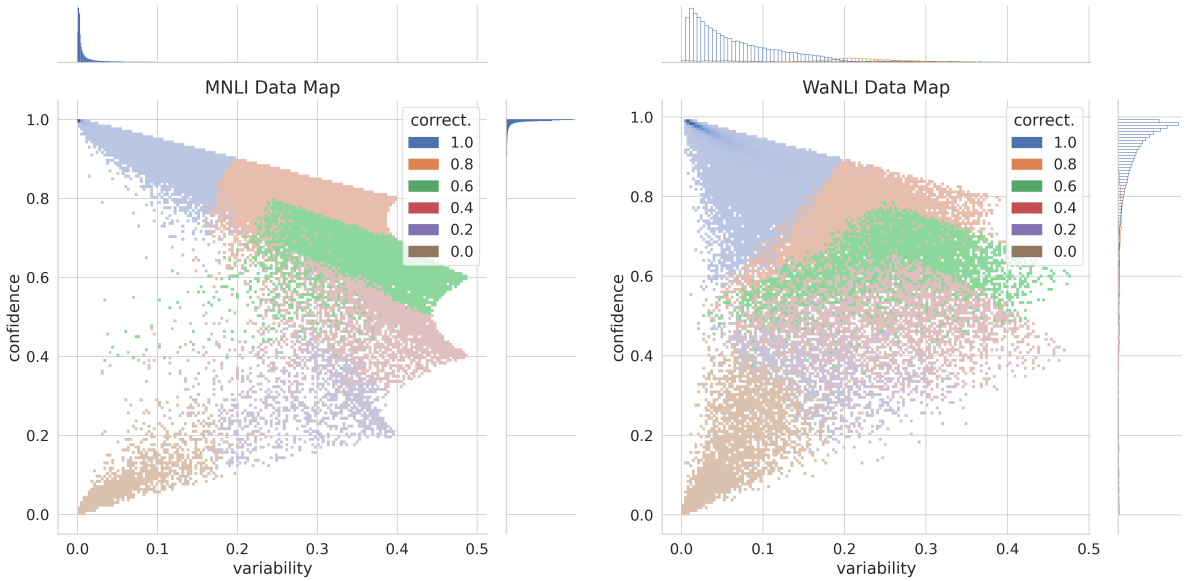

Figure 8: **Left**: Data map for MultiNLI train set, based on a RoBERTa-large classifier trained on MultiNLI. **Right**: Data map for WANLI train set, based on a RoBERTa-large classifier trained on WANLI. A comparison of the distribution in variability (which determines example ambiguity) is remarkable – we see that MNLI 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.

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 7: 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 8: 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 9: 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 10: 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.