Overconfidence in the Face of Ambiguity with Adversarial Data

Margaret Li* **and Julian Michael***

Paul G. Allen School of Computer Science & Engineering, University of Washington
{julianjm,margsli}@cs.washington.edu

Abstract

Adversarial data collection has shown promise as a method for building models which are more robust to the spurious correlations that generally appear in naturalistic data. However, adversarially-collected data may itself be subject to biases, particularly with regard to ambiguous or arguable labeling judgments. Searching for examples where an annotator disagrees with a model might over-sample ambiguous inputs, and filtering the results for high inter-annotator agreement may under-sample them. In either case, training a model on such data may produce predictable and unwanted biases. In this work, we investigate whether models trained on adversarially-collected data are miscalibrated with respect to the ambiguity of their inputs. Using Natural Language Inference models as a testbed, we find no clear difference in accuracy between naturalistically and adversarially trained models, but our model trained only on adversarially-sourced data is considerably more overconfident of its predictions and demonstrates worse calibration, especially on ambiguous inputs. This effect is mitigated, however, when naturalistic and adversarial training data are combined.

1 Introduction

End-to-end neural network models have had widespread success on standard benchmarks in NLP (Wang et al., 2018, 2019; Lee et al., 2017; Dozat and Manning, 2017). However, models trained with maximum-likelihood objectives under the standard Empirical Risk Minimization paradigm are liable to succeed in these settings by fitting to features or correlations in the data which are ultimately not representative of the underlying task and fail to generalize out of distribution, e.g., under domain shift or adversarial perturbation (Gururangan et al., 2018; Ilyas et al., 2019). One promising method to overcome this difficulty is to move past the ERM paradigm and learn or evaluate causal features which are invariant across domains or distributions of data. While methods to do this often require the use of explicitly specified domains of data (Peters et al., 2016; Arjovsky et al., 2020), a more lightweight approach is adversarial evaluation and training (Nie et al., 2020a; Kiela et al., 2021), in which annotators deliberately search for examples on which a model fails. Adversarial data annotation has been applied for a variety of tasks, including question answering (Bartolo et al., 2020), natural language inference (Nie et al., 2020a), hate speech detection (Vidgen et al., 2021), and sentiment analysis (Potts et al., 2021). Adversarial data can help reduce spurious correlations in existing data (Bartolo et al., 2020), expose a model's shortcomings in evaluation, and aid in training more robust models (Wallace et al., 2022).

However, the process of developing adversarial data is imperfect, and adversarial data may itself not resemble naturalistic distributions. For example, Phang et al. (2021) find that the AFLITE adversarial filtering algorithm (Sakaguchi et al., 2020; Bras et al., 2020), designed to find challenging examples in existing datasets, disproportionately favors contentious examples with annotator disagreement. This is suggestive that adversarially *collected* datasets, where humans actively try to fool a model, may be subject to these same biases Indeed, Phang et al. also show that adversarially-collected datasets may disproportionately penalize models that are similar to the one used during data collection. The qualitative properties of adversarially-collected data also vary depending on the adversary used during data collection, as shown by Williams et al. (2022) for the Adversarial NLI dataset (Nie et al., 2020a). For these reasons, it is not clear what a model's performance under adversarial evaluation implies about its performance characteristics on naturalistic distributions, nor is it clear how training on adversarial data aids a model's perfor-

[*] Equal contribution.

Proceedings of the First Workshop on Dynamic Adversarial Data Collection, pages 30 - 40
July 14, 2022 ©2022 Association for Computational Linguistics

mance in natural settings.

In this work, we focus on the interplay of adversarial learning and evaluation with *ambiguity*, or annotator disagreement. Just as adversarial filtering may over-sample ambiguous inputs (Phang et al., 2021), adversarial annotators may produce strange, ambiguous, or disputable inputs as they employ tricks to fool a model in the adversarial setting. To preempt this issue and ensure data quality, adversarial data collection methods filter out examples with low human agreement (Nie et al., 2020a), but it's possible that this approach could over-correct for the issue and *under*-sample such inputs in comparison to naturalistic data. For this reason, it is plausible that models trained on adversarially-collected data may be miscalibrated against the ambiguity of their inputs, forming a predictable blind spot.

We investigate this issue by training models on naturalistically and adversarially collected datasets, then comparing their performance with respect to gold annotator distributions. As a testbed, we use Natural Language Inference, an NLP benchmark task with already-available adversarial data (Nie et al., 2020a) and full annotator distributions (Nie et al., 2020b). We find no clear difference in accuracy between naturalistically and adversarially trained models, but our model trained only on adversarially-sourced data is considerably more overconfident of its predictions and demonstrates worse calibration, especially on ambiguous inputs. On the other hand, including both naturalistic data in training as well — as is standard practice (Nie et al., 2020a) — mitigates these issues. While our results do not raise alarms about standard practices with adversarial data, they suggest that we should keep in mind the importance of including naturalistic data in training regimes moving forward.[1]

2 Background: Robustness and Adversarial Data

Suppose we are interested in learning a conditional probability distribution $p(y \mid x)$. The classical machine learning approach of Empirical Risk Minimization does so with the use of input data drawn from a distribution D:

$$\operatorname*{argmin}_{\theta} \mathbb{E}_{x\sim D, y\sim p(\cdot|x)} - \log p(y|x,\theta), \quad (1)$$

where θ are the model parameters. However, this method can do a poor job of approximating $p(y \mid x)$ when x is drawn from very different distributions than D. One approach which has been used to address this is *robust optimization*, which minimizes the worst-case loss subject to some constraints (Madry et al., 2018; Ghaoui and Lebret, 1997; Wald, 1945). We can view robust optimization as solving a minimax problem:

$$\operatorname*{argmin}_{\theta} \max_{D\in\mathbb{D}} \mathbb{E}_{x\sim D, y\sim p(\cdot|x)} - \log p(y|x,\theta), \quad (2)$$

where $\mathbb{D}$ is a space of possible input distributions, and D is adversarially chosen among them. This formulation invites the question: what if $\mathbb{D}$ includes *all possible distributions?* Then we are free to find *any* x which the model gets wrong, and optimizing the loss effectively should produce a model which is robust to a wide range of distributions and hard to exploit.

This suggests a practical approach to improving robustness which involves actively searching for examples on which a model fails, and using those examples to train new, more robust models. This general approach has been applied in a variety of settings in NLP, such as the Build-It Break-It shared task (Ettinger et al., 2017), adversarial filtering of large datasets (Zellers et al., 2018; Sakaguchi et al., 2020), and adversarial benchmarking and leaderboards (Nie et al., 2020a; Kiela et al., 2021).

One complication that arises when sourcing adversarial data is with ambiguous or arguable examples. Suppose $\hat{\theta}$ perfectly models $p(y \mid x)$. Plugging this into Formula 2 yields $\max_{D\in\mathbb{D}} \mathrm{H}(Y \mid x)$, where D is concentrated on the inputs x which maximize the entropy of Y.

In this context, high entropy in the conditional distribution of Y corresponds to high *annotator disagreement*.[2] When a human searches for an adversarial example, they are looking for a *disagreement* between themselves and the model. In this setting, there may be competition for inclusion in these adversarial tasks between ambiguous examples on

[1] Code to reproduce our experiments is available at `https://github.com/julianmichael/aeae`.

[2] In this work, we assume all annotators implement the same probabilistic labeling function (which we are calling 'gold') and disagreement between annotators arises as an inherent feature of the task we are trying to model. We also assume that approximating annotator behavior on arguable or ambiguous examples is a desirable goal. These are simplifications: in some settings, e.g., the *prescriptive* paradigm of Röttger et al. (2022), we may wish to minimize annotator disagreement to learn a deterministic labeling function. In such settings, model behavior on arguable inputs may be uninteresting from the evaluation perspective, though searching for such examples could be useful for refining the task definition or annotation guidelines. We leave such issues out of scope for this work.

which the model is close to the gold (annotator) distribution and less ambiguous examples where the model is further from gold. Thus an adversarial data generation process may be biased towards input examples which are ambiguous but unhelpful for training.

Formally, a simple way to think about counteracting this may be to explicitly subtract the gold entropy from the loss being minimized:

$$\begin{aligned}\operatorname*{argmin}_{\theta} \max_{D \in \mathbb{D}} \; & \mathbb{E}_{x \sim D, y \sim p(\cdot \mid x)} \\ & - \log p(y \mid x, \theta) + \log p(y \mid x).\end{aligned} \tag{3}$$

Here, the objective focuses the distribution D on examples which maximize the model's KL-Divergence from $p(y \mid x)$, no longer favoring ambiguous examples. Practical approaches to scaling adversarial data collection have applied a similar idea: in Adversarial NLI (Nie et al., 2020a) and Dynabench (Kiela et al., 2021), annotators are asked to find examples where they disagree with the model, and then these examples are only kept if multiple validators agree on the correct label. However, it is not clear how well-calibrated this process is: it might, for example, systematically omit genuinely ambiguous examples which the model gets wrong with high confidence. Whether training on data produced by this process results in pathological model behavior is what we test in this work.

3 Experimental Setup

Task Setting We use Natural Language Inference (Dagan et al., 2005; Bowman et al., 2015) as our underlying task, as there exist adversarial annotations for this task (Nie et al., 2020a; Kiela et al., 2021) and annotator disagreement has been well studied (Pavlick and Kwiatkowski, 2019; Nie et al., 2020b; Zhang and de Marneffe, 2021).

Model Variants We train models under three conditions:

- Classical: These models are trained on data elicited from annotators in a model-agnostic way, i.e., naturalistically.[3] For this we use the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets.

Dataset	Train	Dev
SNLI	550,152	10,000
MultiNLI	392,702	10,000
ANLI (all rounds)	162,865	3,200
Chaos-SNLI		1,514
Chaos-MultiNLI		1,599

Table 1: Number of examples in training and development sets we use. For training data (top), development sets are used for model selection, while our evaluations (bottom) are on the ChaosNLI-annotated subsets of the SNLI and MultiNLI development sets.

- Adversarial: These models are trained on data elicited from annotators under the requirement that they must fool the model. For this we will use the adversarial annotations of Nie et al. (2020a).[4]
- All: These models are trained on the concatenation of all of the above data.

Evaluation Data We test the performance of our models in the setting where we have comprehensive distributions of annotator behavior. For this, we will use the ChaosNLI evaluation sets (Nie et al., 2020b) which have 100 independent annotations for each example (where the task is 3-way multiclass classification). ChaosNLI includes evaluation sets for SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), and αNLI (Bhagavatula et al., 2020, Abductive NLI). Of these, we use the SNLI and MultiNLI sets, since αNLI has a different task format than other NLI datasets. Dataset statistics are shown in Table 1.

Metrics Using densely-annotated evaluation data, we compute several evaluation metrics. Each metric is stratified across non-overlapping ranges of annotator agreement in order to analyze the dependence of model performance (or model differences) on the ambiguity of its input examples. Let $p(y)$ be the empirical distribution of annotator labels for an input example, and $\hat{y}$ be the model's prediction. Then, our metrics are:

- **Accuracy in Expectation:** The expectation of the accuracy of the model against a randomly sampled annotator in ChaosNLI (i.e.,

[3] Unfortunately, since the NLI task is somewhat artificial, there is no "natural" distribution of input texts. This is one of the issues that leads to annotation artifacts in the first place (Gururangan et al., 2018) since some of the input text must be annotator-generated. Regardless, spurious correlations exist in any naturalistic distribution so we will use these training sets as proxies for something naturalistic.

[4] In order for this to properly count as adversarial data for our model, we use the same model family as Nie et al. (2020a), which is BERT-large (Devlin et al., 2019) fine-tuned on SNLI and MultiNLI.

$p(\hat{y})$). We stratify this by the *human accuracy in expectation*, the accuracy of a randomly-sampled human against the plurality vote of all annotators ($\max_y(p(y))$). We use discrete bins to allow for precise comparison of model performance within and between different regimes of ambiguity.

- **Accuracy against Plurality:** The accuracy of the model against the plurality vote of the 100 annotators ($\hat{y} = \max(p(y))$). We also stratify this by human accuracy in expectation.
- **Model perplexity:** The exponentiated entropy of the model's predicted distribution; higher corresponds to more uncertainty. (This is independent of the gold labels.) We stratify this by the perplexity of the human annotator distribution.
- **KL-Divergence:** The KL-Divergence of the model's predicted label distribution against the empirical distribution of annotated labels. This gives a measure of how well-calibrated the model is with respect to the true annotator distribution. We stratify this measure by the entropy of the human annotator distribution.

Accuracy in expectation emulates the typical accuracy computation in an IID empirical risk minimization setting, while accuracy against plurality allows us to measure accuracy scores above human performance (assuming the plurality among 100 annotators can be treated as the ground truth).[5] We also include the annotator distribution as a human reference point (for KL-Divergence, this is 0 by construction).

Implementation Details

In all of our experiments, we begin with RoBERTa-Large (Liu et al., 2019), a masked language model pretrained on a large text corpus comprised of internet and book corpora. We then attach a classifier head and fine-tune each model according to the dataset combinations listed in Section 3. The model was implemented using the AllenNLP library and trained using the AdamW optimizer to maximize accuracy on the combined development sets of the model variant's respective corpora.

4 Results

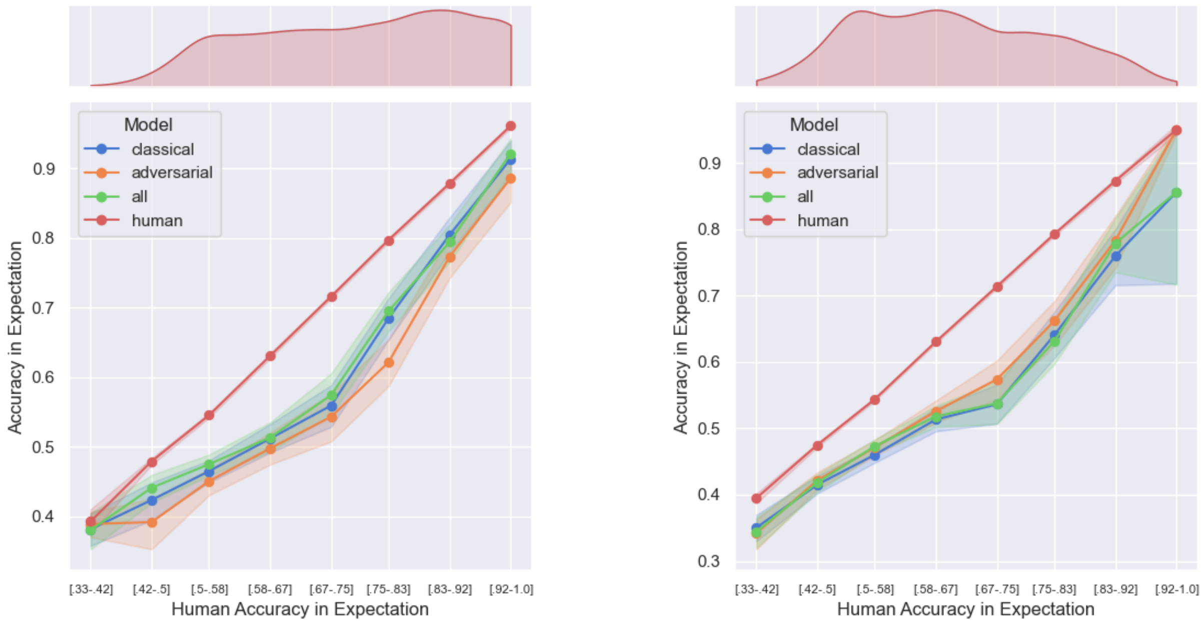
All results in this section are reported on the SNLI and MultiNLI development set portions of the ChaosNLI data. In all graphs, we provide smoothed kernel density estimates of the distributions over X and Y values in the margins where appropriate. Shaded areas around the lines represent 95% confidence intervals.

Accuracy in Expectation Model accuracy against randomly sampled annotators is shown in Figure 1. All models exhibit the same overall trend, approaching or reaching human performance on the most ambiguous and least ambiguous examples, with a dip in the middle of the range. Even if adversarial data collection does under-sample ambiguous inputs, we find no noticeable (or significant) effect on model performance in the low-agreement regime. A potential reason for this is that the baseline performance is already so low in these cases — very close to chance level — that there is little room for decreasing performance further.

Accuracy against Plurality Model accuracy against the plurality vote among annotators is shown in Figure 2. Once again, all models exhibit the same overall trend. While performance seems to level off or even increase for some models on extremely high-ambiguity examples (<50% human accuracy in expectation), there are too few such examples for us to draw any reliable conclusions in this regime.

Perplexity To understand the confidence levels of our models, we measure the perplexity of their output distributions and compare it to the perplexity of the human annotator distributions, shown in Figure 3. Here, there is a clear difference between ADVERSARIAL and the other models: it has extremely low perplexity on many more examples, and high perplexity on very few. Furthermore, while model perplexity is positively correlated with annotator perplexity for all models, the ADVERSARIAL model is less sensitive to it, with its perplexity growing less with respect to annotator perplexity. This suggests the adversarial data collection process may, on aggregate, favor examples with less ambiguity, skewing the behavior of the model. The ALL model, which was exposed to naturalistic data as well, does not display the same effect.

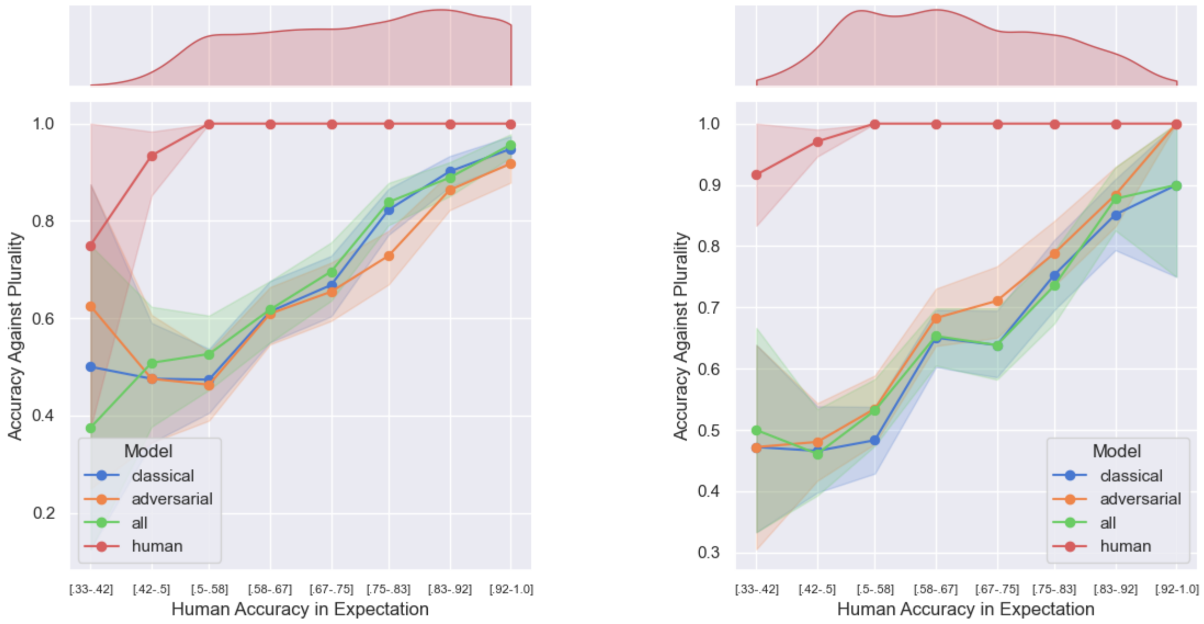
[5] The accuracy metrics provided for NLI datasets in practice are somewhere between the two, as the development and test sets of SNLI and MultiNLI were labeled by 5 annotators each and the majority label was chosen for the purposes of evaluation (Bowman et al., 2015; Williams et al., 2018).



(a) Chaos-SNLI.

(b) Chaos-MultiNLI.

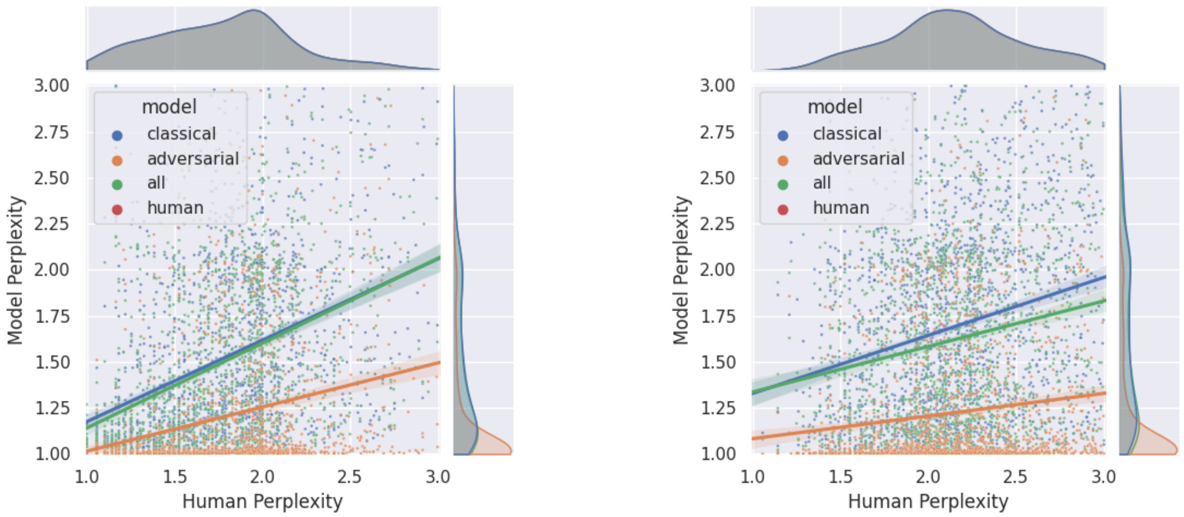
Figure 1: Model accuracy stratified by human accuracy, relative to a randomly sampled human judgment. Chance accuracy is approximately $\frac{1}{3}$, and the human baseline (which uses the plurality vote as the prediction) is an upper bound.



(a) Chaos-SNLI.

(b) Chaos-MultiNLI.

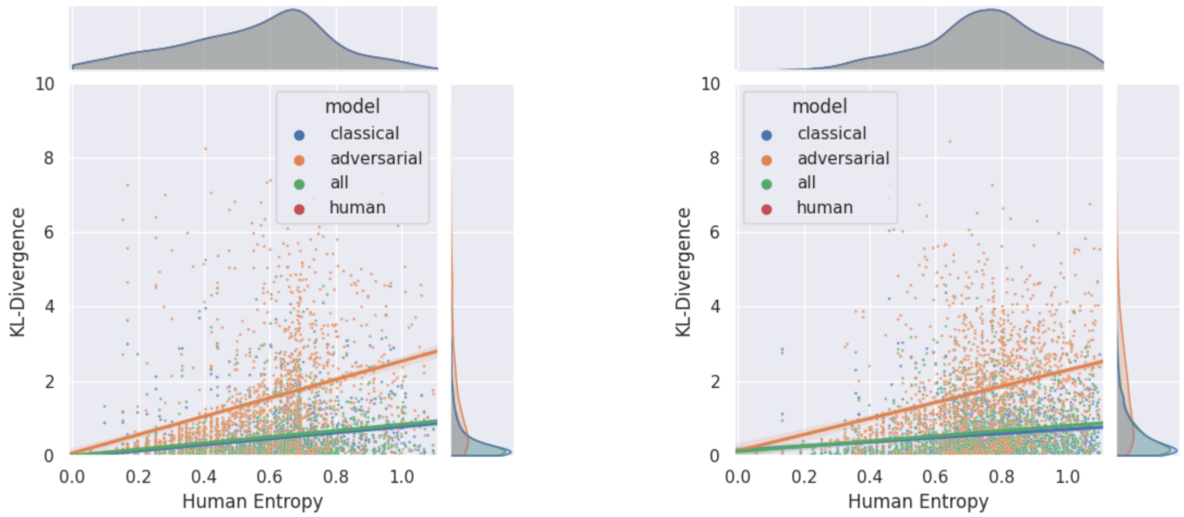
Figure 2: Model accuracy stratified by human accuracy, relative to the human plurality vote. The early dip in the human baseline below 50% is from a few cases with tied plurality votes (where we break ties randomly).



(a) Chaos-SNLI.

(b) Chaos-MultiNLI.

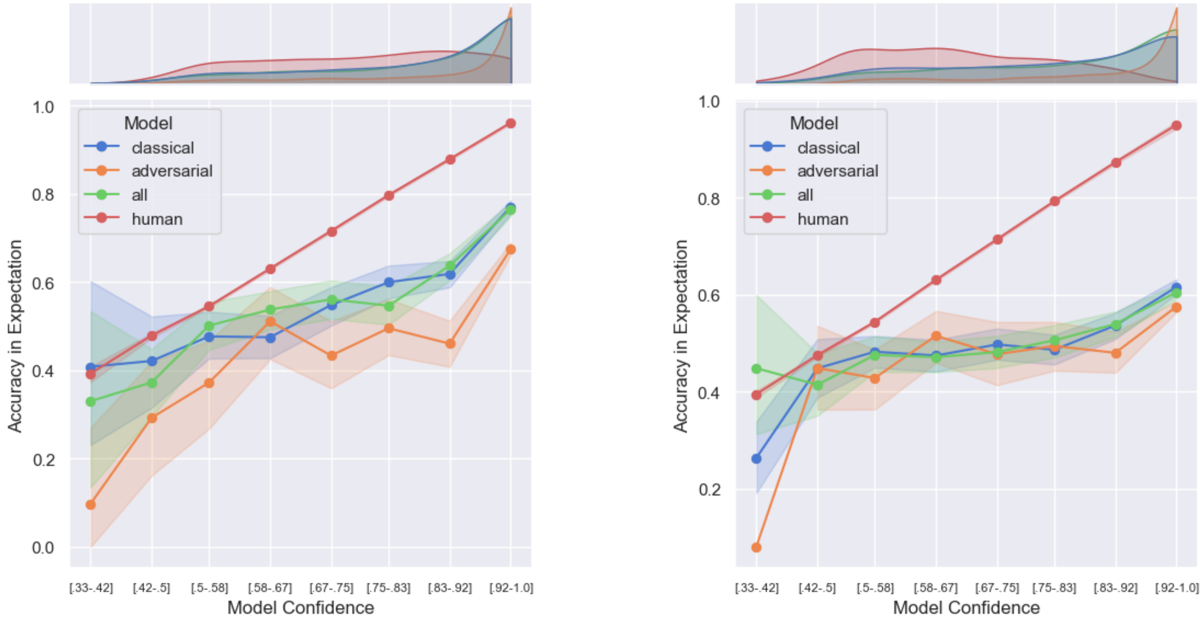
Figure 3: Model perplexity relative to annotator perplexity.



(a) Chaos-SNLI.

(b) Chaos-MultiNLI.

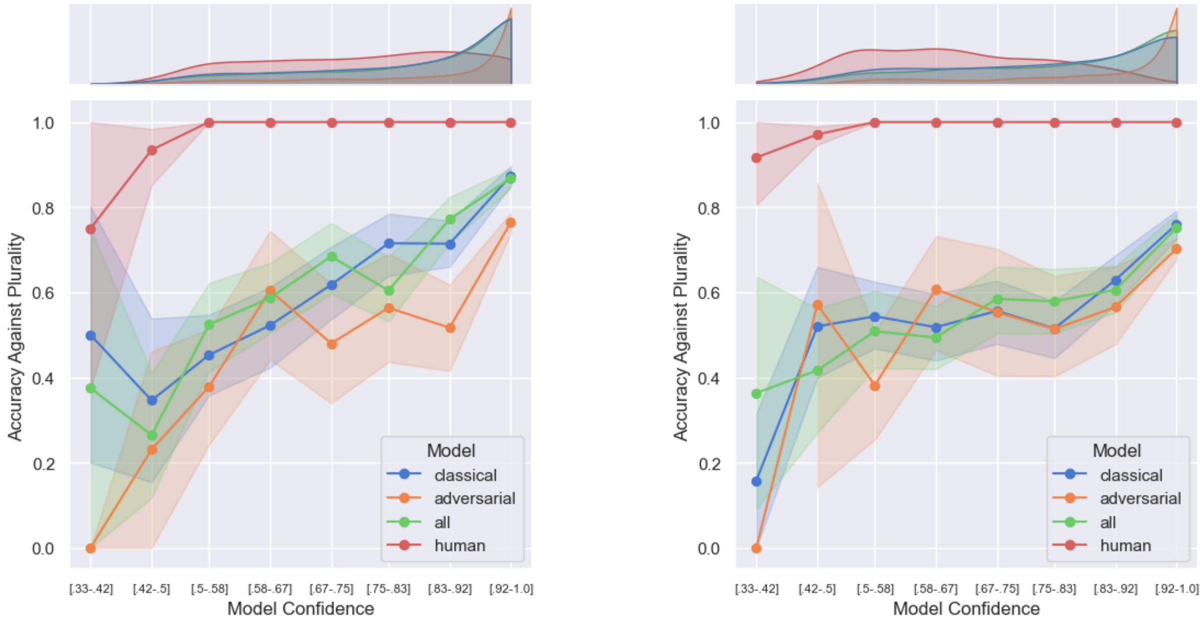
Figure 4: KL-Divergence of model outputs from the annotator distribution, graphed relative to annotator entropy. Both axes are measured in nats.



(a) Chaos-SNLI.

(b) Chaos-MultiNLI.

Figure 5: Calibration curves for accuracy against a randomly sampled human. As the confidence score, we use the probability assigned by the model to its prediction.



(a) Chaos-SNLI.

(b) Chaos-MultiNLI.

Figure 6: Calibration curves for accuracy against the plurality vote among humans. As the confidence score, we use the probability assigned by the model to its prediction.

KL-Divergence To get a sense of how well the model fits the annotator distributions, we show the KL-Divergence of the models' predictions against the annotator distributions in Figure 4. What we find is that ADVERSARIAL diverges greatly from the gold distributions in comparison to CLASSICAL and ALL: it has much higher KL-Divergence in aggregate, many more examples with high KL-Divergence, and its KL-Divergence scores grow more quickly as the entropy of the annotator distribution increases. The biases in adversarial data collection, then, have led more to overconfidence on ambiguous examples than wrong predictions on unambiguous examples. These results provide supporting evidence for the hypothesis that training a model on adversarially-collected data may underexpose it to ambiguous examples and that this could have undesirable effects on its performance. However, these effects seem to be mitigated with the additional inclusion of naturalistic data (in the ALL model).

Calibration Calibration curves are shown in Figure 5. We find that the ADVERSARIAL model is highly confident more often than the other models, and at least in the very-high-confidence regime (>80% confidence), it has significantly worse calibration on SNLI (for MultiNLI, the results are borderline and only for the highest-confidence bin).

We also plot calibration curves relative to the plurality vote among annotators (Figure 6), which reflects the assumption that the model's maximum output probability reflects its epistemic uncertainty over the max-probability label. Here, the results are similar: the ADVERSARIAL model is worse calibrated in the very-high-confidence regime. Note, however, that when optimizing to maximize the likelihood of labels sampled from annotators, the output probabilities of a perfect model will not be well-calibrated against a plurality-based groundtruth. Optimizing for a model calibrated in this way is an alternative design choice which may require different training methods.

5 Discussion

In our experiments using SNLI, MultiNLI, and ANLI, we find that training only on adversarially-collected data produces similar accuracies across all regimes of ambiguity, but worse calibration at high confidence, and more overconfidence on ambiguous examples. This suggests that the adversarial data collection process may bias the model by favoring less ambiguous examples, but there are other potential interpretations of our results.

In particular, the observed miscalibration of ADVERSARIAL may the result of a more general domain shift between SNLI/MultiNLI and ANLI. This could explain why adding SNLI and MultiNLI to training, as in the ALL model, eliminates the effect. However, one might also expect to see a clear difference in accuracy as well if this were the issue. It's also worth noting that the SNLI and MNLI training sets are larger than ANLI's (see Table 1), which could explain why the ALL model behaves similarly to CLASSICAL. It remains an open question how little naturalistic (or, in-domain) data may be sufficient to mitigate the overconfidence issues we observe.

Some notable trends hold for all models we test. First, they all perform worse on ambiguous examples (Figure 1, Figure 2). This may be in part due to the relative scarcity of such examples in the training data or the relative difficulty of learning to model them. Second, they all demonstrate overconfidence, with model perplexity growing slower than human perplexity (Figure 3) and relatively poor calibration at high confidence levels (Figure 6). Even though augmenting training with adversarially-collected data has been shown to improve robustness in some settings (Bartolo et al., 2021a; Vidgen et al., 2021), our results do not yet show any benefits to calibration on ambiguous examples in existing data.

Finally, while we hypothesize that the overconfidence issue with training on adversarial data arises from filtering for annotator agreement, it is also possible that for ANLI, the adversarial annotators found examples that were less ambiguous in the first place (as annotators might, for example, want to focus on sure-fire model mistakes). Williams et al. (2022) found that about 5% of examples in ANLI "could reasonably be given multiple correct labels," suggesting a low level of ambiguity, but this was by the judgment of a single expert and may not correspond to the full variation in label assignment seen with crowdsourced annotators (which could potentially be investigated using the original unfiltered ANLI data). Measuring, controlling, managing, or representing ambiguity in adversarial annotation should be an interesting direction for future work, perhaps incorporating insights from recent work about construal (Trott et al., 2020; Pavlick and Kwiatkowski, 2019), explicit disambiguation (Min et al., 2020), model training dynamics (Swayamdipta et al., 2020; Liu et al., 2022), and other model-in-the-loop adversarial data collection efforts (Bartolo et al., 2020, 2021b; Vidgen et al., 2021; Potts et al., 2021).

6 Conclusion

We have shown that training only on adversarially-collected data, at least in the case of the Adversarial NLI (ANLI) dataset, can produce undesirable performance characteristics in the resulting models. In particular, when tested on SNLI and MultiNLI data, these models produce output distributions that are much further from annotator distributions and fail to accurately convey annotator uncertainty, with highly confident predictions even on highly ambiguous examples. It is also possible that adversarial training in this setting could produce lower prediction accuracy in regimes of low human agreement, but baseline accuracy is already so low for our models and data, and there are so few examples in the extremely-ambiguous regime, that such an effect is hard to find.

In our results, if a large amount of naturalistic data is also included in training (as in the ALL model) — as is standard practice — the overconfidence problem is mitigated. This is encouraging, as any adversarially-collected data must start with some naturalistic data to construct the initial adversary. However, it remains an open question how little naturalistic data is sufficient; a large enough seed corpus may be beneficial for avoiding such issues in a setting of dynamic adversarial data collection (Wallace et al., 2022). Future work can investigate this question, as well as how using full annotator distributions at training time (Zhang et al., 2021) or model calibration techniques may further help models deal with ambiguous inputs.

Acknowledgments

We would like to thank Ludwig Schmidt for early comments on this project, Ofir Press for providing entertainment, and the anonymous reviewers for their useful feedback.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization.

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.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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 (EMNLP)*. Association for Computational Linguistics.

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

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First international conference on Machine Learning Challenges: evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Laurent El Ghaoui and Hervé Lebret. 1997. Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.*, 18(4):1035–1064.

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)*, volume 2, pages 107–112. Association for Computational Linguistics.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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. *CoRR*, abs/2104.14337.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

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

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

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

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Jason Phang, Angelica Chen, William Huang, and Samuel R. Bowman. 2021. Adversarially constructed evaluation sets are more challenging, but may not be fair. *CoRR*, abs/2111.08181.

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

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, WA. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

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.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Abraham Wald. 1945. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 45(2):265–280.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A multi-task benchmark and analysis platform for natural language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3261–3275. Curran Associates, Inc.

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. Association for Computational Linguistics.

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.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54, online. Association for Computational Linguistics.

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

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7620–7632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.