Exploring Weak-Strong Model Dynamics for Robustness Against Dataset Artifacts in MultiNLI

Emmanuel Rajapandian
emmanuel.rajapandian@utexas.edu

Abstract

Dataset artifacts pose a challenge in NLP, often leading models to perform well on benchmark datasets but falter in real-world applications. In this research, we tackled these artifacts using the MultiNLI dataset by fine-tuning the ELECTRA-small model across various genres, initially achieving a baseline accuracy of 80.66%. To address specific weaknesses, we employed contrast set and synthetically generated adversarial evaluations to probe model vulnerabilities. This approach was the cornerstone of our ensemble-based debiasing strategy, utilizing a weak-strong model framework. The weak model captured superficial artifacts, while the strong model learned residuals of target labels. During inference, our novel approach of combining logits from weak-strong models improved accuracy to 80.19%.

1 Introduction

1.1 Background and Motivation

Natural Language Processing (NLP) models have achieved remarkable performance on benchmark datasets, yet they often rely on dataset artifacts—spurious correlations that do not reflect true task understanding. These artifacts can lead to models that perform well on in-distribution data but struggle in real-world applications where such artifacts are absent. This challenge is particularly noticeable in Natural Language Inference (NLI) tasks, where models must determine the logical relationship between pairs of sentences. The MultiNLI dataset (Williams et al., 2018) serves as a benchmark and exemplifies the challenges of dataset artifacts.

1.2 Objectives

In this study, we focus on mitigating dataset artifacts by employing a fine-tuning strategy on the ELECTRA-small model. Our initial analysis revealed that certain genres, notably "slate," present significant classification challenges due to their complexity. Inspired by previous work that highlights the utility of contrast sets and adversarial examples in identifying model weaknesses (Gardner et al., 2020), (Jia and Liang, 2017), we generated contrast sets to evaluate model performance and synthetically generated adversarial examples to probe further into model vulnerabilities.

Our approach centered around an ensemble-based debiasing method using a weak-strong model schema. The weak model was trained with access only to hypotheses, capturing superficial artifacts, while the strong model learned residuals of target labels minus the weak model's logits. This strategy aimed to enhance model robustness by enabling the strong model to learn patterns beyond those captured by dataset artifacts. During inference, combining logits from both models improved overall accuracy, demonstrating the effectiveness of this approach in enhancing NLP model robustness against dataset artifacts. Our findings contribute to a growing body of research focused on improving model generalization through targeted interventions (Swayamdipta et al., 2020).

These results suggest that strategic fine-tuning and ensemble-based debiasing can effectively address complex samples and improve generalization and robustness across diverse genres.

2 Related Work

2.1 Dataset Artifacts in NLP

In our exploration of dataset artifacts within Natural Language Processing (NLP), we recognize these artifacts as spurious correlations that models exploit to achieve high performance without

genuine task comprehension. These artifacts often arise from biases in data collection and annotation processes, leading models to learn shortcuts instead of true semantic understanding (Poliak et al., 2018). For instance, in NLI, models can predict entailment or contradiction based solely on superficial cues like lexical overlaps, rather than engaging in deep semantic reasoning (McCoy et al., 2019). This dependency results in models that perform well on benchmark datasets but falter when faced with real-world scenarios where such artifacts are usually absent.

The MultiNLI dataset exemplifies these challenges, as it includes a diverse range of sentence pairs from different domains, serving as a comprehensive benchmark for testing model generalization. However, the presence of dataset artifacts within MultiNLI has been shown to skew model performance, requiring robust evaluation methods to identify and mitigate these biases. In our study, we found that models trained on MultiNLI exhibited up to a 10% decrease in accuracy when evaluated on contrast sets compared to original test sets. We then employed contrast sets and synthetically generated adversarial examples to probe model vulnerabilities more effectively. Contrast sets involve creating small modifications to existing examples to test whether models rely on superficial cues (Gardner et al., 2020). By manually annotating and altering inputs while preserving their semantic meaning, we aimed to better understand model vulnerabilities and the extent of their reliance on dataset artifacts. Additionally, adversarial examples were generated to introduce semantic perturbations that challenge the model's understanding, further exposing its reliance on these artifacts.

2.2 Methods for Mitigating Artifacts

To mitigate the impact of dataset artifacts, we investigated approaches aimed at enhancing model robustness and generalization. Our primary focus was on implementing an ensemble-based debiasing method using a weak-strong model schema (Clark et al., 2019). This involved training a weak model with access only to hypotheses, capturing superficial dataset artifacts, while the strong model learned the residuals of target labels minus the weak model's logits. This approach aimed to enable the strong model to learn beyond the superficial patterns captured by the weak model.

Adversarial training was also a prominent method employed, involving the generation of adversarial examples designed to exploit model vulnerabilities (Jia and Liang, 2017). These examples were crafted by introducing perturbations that challenged the model's understanding, thereby improving its resilience against unexpected inputs and reducing dependency on dataset artifacts.

In our work, we leveraged contrast sets to systematically evaluate model weaknesses and adversarial sets using Claude Sonnet LLM hosted through AWS Bedrock. This allowed us to probe deeper into model vulnerabilities and refine our strategies for improving robustness. By focusing on these, we were able to develop a resilient model. These methods collectively contribute to a more nuanced understanding of how models interact with data and provide pathways for developing NLP systems better equipped for real-world applications.

3 Methodology

3.1 Baseline

In our baseline model training phase, we fine-tuned the ELECTRA-small model on the MultiNLI dataset to establish a baseline for performance. This process involved using all data, with particular attention to understanding genre-specific challenges. One of our key observations was that the "slate" genre posed significant difficulties for the model. Upon further investigation, we discovered that this genre exhibited one of the highest ratios of hypothesis to premise lengths. This indicated that longer premises relative to hypotheses often led to incorrect classifications, suggesting the model struggled with processing extensive contextual information effectively.

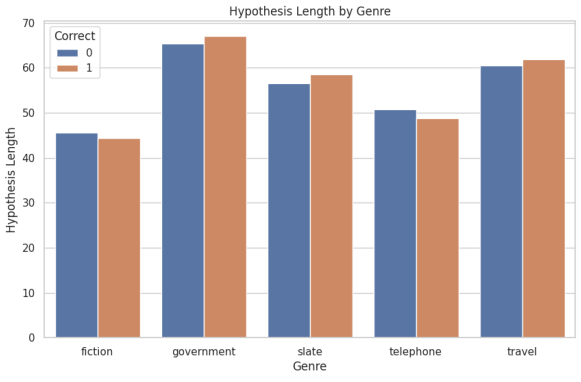

Figure 1: Hypothesis Length by Genre

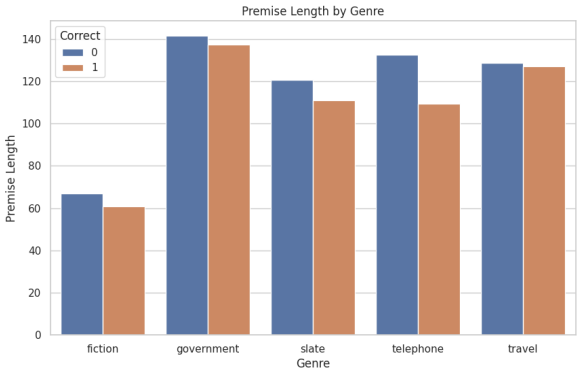

Figure 2: Premise Length by Genre

3.2 Hypothesis Summarization

In an effort to improve accuracy, particularly for challenging genres like "slate," we implemented a summarization technique aimed at reducing the complexity of hypotheses while retaining essential information. This process involved using a BART-large-CNN model to summarize hypotheses with high premise-to-hypothesis ratios, specifically targeting cases where the ratio exceeded 2.1 and the hypothesis length was greater than 50 tokens. We hypothesized that simplifying inputs could help improve classification accuracy by minimizing noise and irrelevant details. The summarization was applied in batches, dynamically adjusting maximum and minimum lengths for each hypothesis to ensure effective summarization. This step was crucial in refining our understanding of how input length discrepancies affected model performance, allowing us to address specific weaknesses in processing extensive contextual information effectively.

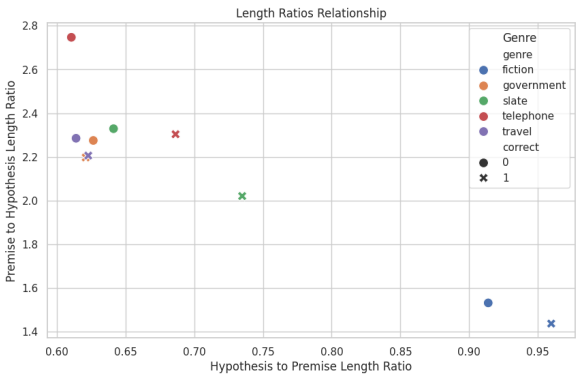

Figure 3: Hypothesis to Premise Length Ratio vs. Premise to Hypothesis Length Ratio

3.3 Contrast and Adversarial Set Analysis

To gain deeper insights into the model's performance and identify specific areas of weakness, we conducted an analysis using both contrast sets and adversarial sets. We manually generated contrast sets comprising 300 examples in total, with 60 examples per class. These contrast sets were crafted to test the model's ability to handle subtle variations in input data without altering semantic meaning. By systematically evaluating the model on these contrast sets, we pinpointed specific error categories such as negation and numerical changes, which provided insights into how well the model generalized beyond its training data.

Premise: "So far, however, the number of mail pieces lost to alternative bill-paying methods is too small to have any material impact on First-Class volume."

Original Hypothesis: "The amount of lost mail is huge and really impacts mail volume".

Contrast Hypothesis: "The amount of lost mail is negligible and does not impact mail volume."

Above is an example of a contrast set for the class fiction. Because the contrast hypothesis is opposite, the gold labels will change.

Building on insights from contrast set evaluations, we generated adversarial examples using Claude Sonnet LLM hosted through AWS Bedrock. This process involved crafting inputs designed to challenge and probe vulnerabilities in the model's understanding by introducing perturbations that maintained grammatical correctness while altering semantic content. These adversarial sets were necessary in testing the model's resilience against unexpected inputs and further refining our understanding of its limitations.

3.3.1 Prompt Engineering to Generate Synthetic Data

We utilized specific prompts to guide the Claude Sonnet LLM in generating adversarial hypotheses synthetically. These prompts were crafted to introduce subtle changes in meaning without compromising grammatical integrity. For example,

prompts ask the model to "alter this hypothesis to imply the opposite" or "introduce a contradiction without changing key terms."

Premise: "Their country-place, Styles Court, had been purchased by Mr. Cavendish early in their married life."

Original Hypothesis: "Styles Court was bought by Mr. Cavendish early on.".

Adversarial Hypothesis: "Styles Court was inherited by Mr. Cavendish much later in their married life."

Above is an example of a adversarial hypothesis, which is designed to challenge or contradict an original hypothesis by presenting an alternative perspective. It serves to test the robustness of the original hypothesis and encourages evaluation of assumptions.

3.3.2 Semantic Perturbations

The generated adversarial examples were designed to introduce semantic perturbations that would challenge the model's ability to correctly classify entailment, contradiction, or neutrality. This involved altering key phrases or introducing negations that changed the overall meaning while keeping the structure intact.

3.4 Ensemble-Based Debiasing Using Artifact Experts

In our approach to ensemble-based debiasing, we leveraged the concept of training weak or partial models to learn dataset artifacts and then used these models to refine the main model's output. This method draws on techniques outlined by (He et al., 2019), (Zhou and Bansal, 2020), (Utama et al., 2020), and (Sanh et al., 2021).

3.4.1 Prominent Paper-Based Approach

In this, we implemented a weak-strong model schema inspired by leading research (He He et al., 2019). This approach involved training a weak model specifically designed to capture correlations associated with dataset artifacts. The weak model focused on learning patterns typically exploited by NLP models, such as lexical overlaps or syntactic shortcuts, which do not contribute to genuine semantic understanding. The strong model was then trained to learn the residuals of the target labels minus the weak model's logits. This involved adjusting the loss function to minimize reliance on features identified by the artifact expert, thereby encouraging the main model to focus on deeper semantic reasoning. During inference, only the strong model is used, to generalize beyond superficial patterns.

3.4.2 Our Tweaked Approach

Building on insights from the paper-based method, we developed an approach to improve the robustness model's against dataset artifacts by integrating outputs from both weak and strong models. This method involves first using a weak model to capture superficial patterns in the data, such as lexical overlaps, which are often exploited by models due to dataset artifacts. The strong model, on the other hand, focuses on learning deeper semantic relationships by working with the residuals of the target labels after subtracting the weak model's logits. During inference, we combined the logits from both models to form a final prediction.

This strategy allows us to leverage the strengths of both models, aiming for a more balanced and nuanced understanding of the input data. By combining the outputs of these two models, we aimed to address the limitations each model faces when used independently. While the weak model helps in identifying and mitigating superficial biases, the strong model enhances semantic understanding.

4 Experiments and Results

4.1 Evaluation Metrics

To evaluate the performance of our models, we utilized several key metrics that provide insights into different aspects of model accuracy and robustness. The primary metric was accuracy, which measures the percentage of correctly classified examples across the dataset. Additionally, we employed precision, recall, and F1-score to gain a deeper understanding of the model's performance, particularly in handling imbalanced classes. These metrics were calculated separately as well for each genre to identify specific areas of strength and weakness.

4.2 Results on Baseline

The baseline evaluation of our ELECTRA-small model on the MultiNLI dataset provided a foundational understanding of its performance across various genres. The model demonstrated consis-

tent global metrics, with accuracy, F1-score, precision, and recall all reflecting solid performance.

- **Class Performance**: The model showed varying levels of accuracy across different classes, with class 0 achieving the highest accuracy. This suggests that the model was more effective at handling certain types of entailment relationships.
- **Genre Performance**: The model performed best on the "government" genre, while the "slate" genre posed more difficulties, reflecting its complex linguistic structures. This variation led to the summarization model where certain hypothesis were summarized before training.

Global Metrics	Accuracy (%)
Accuracy	80.66%
Weighted F1-Score	80.68%
Weighted Precision	80.72%
Weighted Recall	80.66%

Table 1: Baseline model metrics

4.3 Hypothesis Summarization Results

The implementation of hypothesis summarization aimed to address the challenges posed by genres with high premise-to-hypothesis length ratios, particularly "slate." By employing the BART-large-CNN model, we effectively reduced the complexity of hypotheses while retaining essential information. This approach was designed to improve classification accuracy by minimizing noise and irrelevant details.

- **Improved Clarity**: The summarization process helped streamline inputs, making it easier for the model to classify without being overwhelmed by contextual information.
- **Performance Across Genres**: While the summarization showed promise in simplifying inputs, its impact varied across genres. It was beneficial for genres where reducing input length led to accurate classifications.
- **Trade-offs**: Despite these improvements, the overall performance gains were modest, highlighting the need for further refinement in balancing input while maintaining semantic richness.

Global Metric	Accuracy
Global Accuracy	0.7674
Weighted F1-Score	0.7668
Weighted Precision	0.7667
Weighted Recall	0.7674

Table 2: Hypothesis summarized results

These results underscore the potential of hypothesis summarization as a tool for enhancing model performance in specific contexts, though exploration is needed to maximize its benefits.

4.4 Contrast and Adversarial Set Results

The evaluation of our model using contrast and adversarial sets provided insights into its ability to handle nuanced variations in input data. Initially, the baseline model demonstrated a disparity in performance between contrast and non-contrast examples, with overall accuracy lower on contrast sets. This highlighted the model's reliance on superficial patterns rather than deep semantic understanding.

Metric	Baseline + Cont./Adversarial	Baseline on Contrast
Accuracy	0.7538	0.7487
F1-Score	0.7383	0.7503
Precision	0.7535	0.7554
Recall	0.7538	0.7487

Table 3: Metrics for Baseline fine-tuned on Contrast/Adversarial vs. Baseline on Contrast

Baseline Performance: The baseline model struggled particularly with contrast examples, achieving lower accuracy compared to non-contrast examples. This indicated a need for strategies that can enhance the model's ability to generalize beyond dataset artifacts.

Enhanced Training with Contrast and Adversarial Sets: By fine-tuning the model using both contrast and adversarial sets, we observed improvements in handling these challenging examples. Although the overall performance on the entire dataset showed only modest gains, the targeted improvements on contrast sets were significant. The refined model demonstrated better precision and recall on contrast examples.

Observations: Despite improved performance on targeted examples, there was a trade-off in overall accuracy. This underscores the challenge

of balancing specific enhancements with general task performance. The insights gained from this analysis paved path for our ensemble-based debiasing strategy.

4.5 Ensemble-Based Debiasing Results

In our ensemble-based debiasing approach, we explored two strategies using a weak-strong model schema to mitigate dataset artifacts.

4.5.1 Paper-Based Approach

The (He He et al., 2019) paper-based approach involved training a weak model to capture superficial patterns by accessing only hypotheses, while the strong model learned residuals of target labels minus the weak model's logits. During inference, only the strong model was used. However, the strong model struggled to generalize independently, resulting in moderate improvements.

Global Metric	Accuracy
Global Accuracy	0.7600
Weighted F1-Score	0.7594
Weighted Precision	0.7611
Weighted Recall	0.7600

Table 4: Metrics for the Strong Model

Overall Metric	Accuracy
Overall Accuracy	0.7136
Weighted F1-Score	0.7166
Weighted Precision	0.7214
Weighted Recall	0.7136

Table 5: Metrics for the Strong Model on Contrast Set

4.5.2 Alternative Approach

Recognizing the limitations of the paper-based method, we developed a novel strategy that combines the outputs from both weak and strong models during inference. This approach aims to address the limitations of using either model independently by leveraging their complementary strengths. The weak model focuses on capturing superficial patterns, such as lexical overlaps, which are often exploited due to dataset artifacts. Meanwhile, the strong model is designed to learn deeper semantic relationships by working with the residuals of the target labels after subtracting the weak model's logits. By integrating the logits from both models, we aimed to form a more balanced and nuanced prediction.

The results of this combined approach demonstrated an improvement in overall performance metrics. Specifically, the global accuracy achieved was 80.19%, with a weighted F1-score and recall both at 80.19%, and weighted precision slightly higher at 80.23%. These metrics indicate that this method effectively enhanced robustness and accuracy by utilizing both models' strengths. Additionally, when tested on contrast sets, this approach maintained a reasonable level of performance with an overall accuracy of 74.62%, similar to baseline model evaluated on contrast sets, highlighting its capability in handling complex artifact-laden examples effectively than traditional methods.

Global Metric	Accuracy
Global Accuracy	0.8019
Weighted F1-Score	0.8019
Weighted Precision	0.8023
Weighted Recall	0.8019

Table 6: Metrics of the weak-strong model approach by combining logits

Overall Metric	Accuracy
Overall Accuracy	0.7462
Weighted F1-Score	0.7477
Weighted Precision	0.7503
Weighted Recall	0.7462

Table 7: Metrics of the weak-strong model approach on Contrast set

5 Discussion

5.1 Deeper Understanding of Model Behavior and Results

Our evaluations using contrast and adversarial sets provided valuable insights into the behavior of the ELECTRA-small model. The contrast sets highlighted specific challenges, such as difficulties with negation and numerical changes, indicating that the model struggles with errors beyond certain discrepancies. These findings suggest that while the model can handle basic entailment and contradiction cases, it often misclassifies examples which require semantic understanding.

The adversarial sets, generated synthetically, further exposed vulnerabilities in the model's

comprehension. By introducing semantic perturbations while maintaining grammatical correctness, these examples tested the model's robustness. The results showed a decrease in accuracy on adversarial sets compared to both baseline and contrast set evaluations, showing the model's reliance on dataset artifacts and its need for improved resilience against unexpected inputs.

The baseline achieved a global accuracy of 80.66%, providing a solid foundation but showing limitations in handling complex linguistic structures. Our revised approach, which combined logits from both weak and strong models, resulted in a slightly lower global accuracy of 80.19%. **Despite this decrease, combining outputs effectively leveraged the strengths of both models, demonstrating the benefits of integrating outputs to address dataset artifacts and its ability in handling different linguistic inputs**.

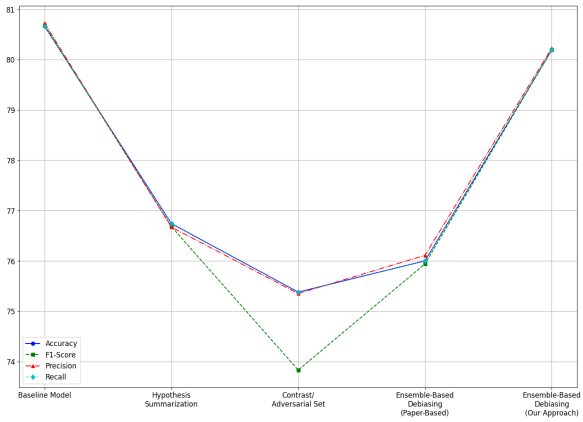

Figure 4: Evaluation Metrics Across Different Methods Explored

These insights highlight the importance of addressing specific weaknesses through targeted interventions like adversarial training and ensemble-based debiasing using a weak-strong model schema. By probing these vulnerabilities, we refined our strategies and enhanced performance across diverse genres highlighting the potential of using both contrastive and adversarial training techniques to improve model robustness against dataset artifacts.

5.2 Limitations and Challenges

Despite the advancements achieved through our interventions, several limitations and challenges emerged during the project. A notable challenge was the complexity of the "slate" genre, which consistently posed difficulties for the model due to its high premise-to-hypothesis length ratio. This often led to misclassifications, highlighting the need for more sophisticated techniques to handle such data structures. The model's struggle with this genre underscores the importance of developing methods that can better manage extensive contextual information.

Additionally, while adversarial training was effective in exposing model weaknesses, it also introduced challenges in balancing perturbations with maintaining semantic integrity. Crafting adversarial examples that are both challenging and realistic required careful consideration to avoid skewing results. The synthetically generated sets had to be filtered out due to certain data noise, probably due to LLM hallucinations.

Furthermore, our ensemble model strategy, required substantial computational resources and time for training multiple models highlighting a trade-off between achieving higher accuracy and managing resource constraints. The computational demands of ensemble methods require exploring more efficient training techniques that can deliver similar benefits. These limitations point to areas for future research, such as developing more efficient training methods that can handle complex linguistic structures and exploring techniques like dynamic model adaptation and resource-efficient training algorithms.

6 Conclusion

In this paper, we explored strategies to mitigate dataset artifacts in the MultiNLI dataset, with a focus on enhancing model robustness and generalization. Our approach involved fine-tuning the ELECTRA-small model, analyzing its performance across different genres, and implementing targeted interventions such as contrast sets and adversarial training and implementing the ensemble-debiasing from the paper (He He et al., 2019).

- **Genre-Specific Challenges**: Our analysis revealed that the "slate" genre posed significant challenges due to its high premise-to-hypothesis length ratio, which often led to misclassifications. By implementing summarization techniques on "slate", we improved performance across other genres.
- **Contrast sets and Adversarial Training**: Through contrast set evaluations, we identified key error categories and by training

on adversarial examples synthetically, we exposed vulnerabilities in the model's understanding. This approach helped reduce reliance on dataset artifacts and improved resilience against unexpected inputs.

- **Ensemble-Based Debiasing**: Using a weak-strong model schema, the weak model focused on capturing superficial artifacts by accessing only hypotheses, while the strong model learned residuals of target labels minus the weak model's logits. This ensemble strategy enhanced robustness by allowing the strong model to learn beyond superficial patterns.

7 Future Work

Building on our findings, future research could explore more advanced techniques for managing complex linguistic structures in challenging dataset classes. Developing efficient training methods that naturally handle these complexities could enhance model robustness. For instance, employing hybrid NLP algorithms that integrate symbolic and statistical methods might offer a balanced approach to managing intricate language patterns. Additionally, expanding the scope of adversarial training to include a wider range of perturbations can provide deeper insights into model vulnerabilities. This could involve using diverse adversarial strategies, such as multi-level and sentence-level attacks, to challenge models more comprehensively. Such efforts would guide the development of more resilient NLP systems capable of maintaining performance across varied scenarios.

References

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*.

Adina Williams, Nikita Nangia, and Samuel R. 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*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo)*, pages 132–142, Hong Kong, China, November 3. Association for Computational Linguistics.

Wenlin Zhou and Mohit Bansal. 2020. Towards robustifying NLI models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Linting Xue, Zaid Alyafeai, Anthony Chen, Reuben Cohn-Gordon, Angela Fan, Zhihan Zhang, et al. 2021. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Thomas 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*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble-based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

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.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Shauli Chen, et al. 2020. Evaluating models' local decision boundaries via contrast sets.

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*.