UNIVERSITÄT
TRIER

Department of Computational Linguistics and Digital Humanities

Natural Language Processing

Unlearning Shortcuts: Targeted Debiasing for Improved Inference

Sanjana Deshpande

Matriculation Number:1763209

Lecturer: Dr Simon Werner

January 11, 2026

Declaration

I, Sanjana Deshpande, hereby declare that:

- this work is my own and original. As author, I am responsible for the provenance and veracity of all statements and claims made in this work.
- parts of this work that were taken from other works, either as quote or paraphrase, either direct or indirect, or obtained with the help of a third party are marked by respective statements of sources.
- parts that were generated or obtained using tools (like generative AI or other software) that influenced the content of this work are also marked. Such tools are transparently listed in this paper. Specifics about the usage of those tools (how they have been used and with what effect) is being reported.
- this work has not been handed in for a different academic assessment by me or another person.

Signature

Sanjana Deshpande
January 11, 2026

Contents

Chapter 1

Introduction

1.1 Background

Empirical work on Natural Language Inference has shown that high performance is not indicative of inference based genuine semantic understanding. Standard datasets contain annotation artifacts that models exploit as shortcuts, leading to systematic failures on controlled perturbations. These failure modes are well-documented under error taxonomies such as negation bias, syntactic insensitivity, etc. These findings suggest that current models optimize for dataset heuristics rather than robust reasoning.

1.2 Problem statement

The persistence of such biases highlights a fundamental limitation: models trained on standard datasets optimize for patterns and correlations rather than semantic reasoning. This produces brittleness in evaluation, specially significant drops in performance on test sets that explicitly probe known error categories. This leaves an open question as to whether dataset-level debiasing can meaningfully improve generalization in these categories. The outcome is twofold: debiasing may either help models unlearn superficial cues, or alternatively reduce their ability to exploit dataset regularities, lowering overall performance. A careful empirical evaluation is therefore required.

1.3 Aims and objectives

The objectives are deliberately diagnostic rather than purely performance-driven:

- Assess whether debiased training improves robustness on a variety of known error categories.
- Quantify the trade-off between overall accuracy and targeted robustness, determining whether debiasing sacrifices global performance for local gains or yields improvements on both fronts.
- Characterize how error categories react to debiasing, thereby shedding light on whether bias removal primarily addresses lexical-level artifacts or extends to structural perturbations.
- Conduct fine-grained error analysis to identify the qualitative nature of residual failures.

1.4 Solution approach

The study adopts a comparative experimental framework to assess the impact of debiasing on NLI robustness on a model against its baseline. Two models are fine-tuned separately: one on a standard training corpus and another on a version of the same corpus processed to reduce dataset biases. Both models are then evaluated on a controlled test set with systematically augmented examples targeting known error categories, such as negation, lexical overlap, inversion, and passivization. Performance is analyzed along three dimensions: overall improvement in inference, accuracy within individual error categories, and qualitative error inspection. This setup allows for a focused investigation of whether debiasing reduces systematic vulnerabilities, as opposed to merely boosting aggregate benchmark performance.

1.5 Summary of contributions and achievements

I successfully brought to light the systematic vulnerabilities of NLI models caused by dataset-induced biases and demonstrated that models can reduce reliance on superficial cues.

1.6 Organization of the report

This report presents an analysis of whether debiasing training data improves the generalization of NLI models across known error taxonomies, following a structure that moves from conceptual foundations to empirical evaluation, and finally to broader implications.

- **Chapter 2: Technical Context and Related Work** – This chapter reviews prior research on biases in NLI, including error taxonomies such as negation and lexical overlap. It also discusses prior approaches to debiasing and data augmentation, positioning this study in the broader effort to improve inference in models.
- **Chapter 3: Methodology** – This chapter details the experimental framework, explaining the use of debiasing in fine-tuning, the design of controlled test-time augmentations for evaluating specific error categories, and the metrics used for assessing model performance.
- **Chapter 4: Experimentation and Results** – This chapter outlines the experimental setup and presents results comparing the baseline and debiased models. It presents the results in a top-down approach from overall improvements in inference to qualitative insights obtained from manual inspection.
- **Chapter 5: Conclusions** – The report concludes by synthesizing the quantitative and qualitative insights, discussing the extent to which debiasing impacted NLI task.

Chapter 2

Technical Context and Related Work

2.1 Bias in NLI

Gururangan et al. (2018) demonstrated that widely used NLI datasets, such as SNLI and MNLI, contain annotation artifacts that allow models to predict the correct label based solely on hypothesis cues. Poliak et al. (2018) extended this observation, showing that hypothesis-only models, without any access to the premise, perform surprisingly well. These findings established that many benchmark results are inflated by superficial correlations rather than robust inference.

2.2 Error Taxonomies in NLI

Naik et al. (2018) introduced stress tests to evaluate model robustness to phenomena such as negation, antonymy, numerical reasoning, and word overlap. Their work revealed that models frequently misclassify hypotheses with negation cues as contradictions, even when the gold label is entailment or neutral. Glockner et al. (2018) demonstrated that simple lexical substitutions could drastically reduce model accuracy, highlighting models' reliance on superficial lexical heuristics. Together, these studies provide a framework for error taxonomies, identifying predictable categories of model failure.

2.3 Debiasing Approaches

He et al. (2019) introduced residual fitting, a method that explicitly models the bias in the training data and trains models to predict the residual, thereby reducing reliance on dataset-specific artifacts. Belinkov et al. (2019) examined ensemble-based debiasing, combining multiple predictors to prevent overfitting to known heuristics, demonstrating improved robustness. Utama et al. (2020) proposed a self-debiasing framework that regularizes the model during training, encouraging it to avoid overconfident predictions on biased examples and improving generalization without explicit knowledge of the biases.

Complementing model-based debiasing, Wu et al. (2022) introduced a data-generation framework to mitigate spurious correlations in NLI datasets. Their method trains generative models to produce label-consistent examples and applies a filtering mechanism to remove samples that contribute to dataset biases. Evaluation on SNLI and MNLI showed that models trained on these debiased datasets generalized better to out-of-distribution and adversarial sets. Wu et al.'s work highlighted that modifying the training data itself can be an effective route to reducing bias.

2.4 Data Augmentation

Min et al. (2020) showed that NLI models often fail on syntactically challenging examples due to reliance on superficial heuristics. They introduced syntactic data augmentation applied during training, generating new examples using inversion and passivization of hypotheses. Incorporating these augmented examples improved model robustness to structural variations, demonstrating that targeted syntactic transformations can expose models to constructions they would otherwise struggle to handle.

Nguyen (2023) identified significant vocabulary biases in NLI datasets, such as SNLI and MNLI, where certain words are disproportionately associated with specific entailment classes. To mitigate these biases, the study introduced automated data augmentation techniques that operate at each level. Character-level transformations included random insertion and deletion of characters, word-level augmentations involved synonym replacement and random word swaps, and sentence-level modifications encompassed paraphrasing and controlled perturbations. By incorporating these augmented examples during training, Nguyen demonstrated that models achieved improved accuracy and reduced bias, outperforming baseline models.

2.5 Positioning of This Work

This work positions itself at the intersection of dataset-level debiasing and test dataset augmentation for evaluation, investigating whether these interventions improve performance on targeted error taxonomies: negation, lexical overlap, and structural perturbations.

Chapter 3

Methodology

3.1 Overview

The experimental design is structured to isolate the effect of debiasing on model robustness in NLI. To this end, two fine tuning conditions are established: one reflecting a standard learning regime by using the original MNLI dataset (Williams et al., 2018) and the other incorporating debiasing by using a version of MNLI intended to reduce reliance on dataset artifacts. By holding all other factors constant, the comparison focuses exclusively on the contribution of debiasing.

Evaluation is performed on a specially constructed test environment. The evaluation corpus integrates systematically augmented samples that instantiate well-documented error taxonomies. These augmentations cover phenomena such as negation, lexical overlap, syntactic inversion, passivization and synonym transformations. The goal is to directly probe categories where prior work has shown models to be particularly brittle.

This methodology ensures that the evaluation does not simply report improved benchmark scores, but instead rigorously tests whether debiasing translates into more reliable generalization across the linguistic phenomena that have historically exposed the limitations of NLI models.

3.2 Debiasing Approach

This work leverages the debiasing framework of Wu et al. (2022) which integrates data generation with bias filtering. The method incorporates a filtering mechanism to remove generated examples that strongly correlate with known dataset biases. This involves identifying surface-level features, such as frequent hypothesis constructions, negation words, or other patterns statistically associated with particular labels and eliminating or replacing examples that rely on these cues. The resulting debiased training set maintains coverage across all NLI labels (entailment, neutral, contradiction) but encourages models to focus on semantic reasoning rather than memorizing artifacts. Models fine-tuned on this debiased dataset are expected to be less prone to errors arising from lexical overlap, negation, or other superficial correlations. For baseline comparison, a separate model is fine-tuned on the original MNLI dataset, allowing direct assessment of the impact of debiasing on systematic errors.

3.3 Augmentation Procedure

To construct the evaluation set, systematic augmentation techniques inspired by Min et al. (2020), Sullivan (2024) and Nguyen (2023) are applied to a reserved portion of NLI data. In this study, these augmentation strategies are applied purely to the test set to simulate controlled error categories. Negation examples transform hypothesis by prefixing "It is not true that", and lexical overlap examples involve high word overlap between premise and hypothesis that may mislead the model into predicting entailment despite contradictory or neutral meaning. Structural perturbations involve inversion examples that swap subject and object roles, passivization examples that convert sentences between active and passive forms, and synonym examples that replace words in the hypothesis with their synonyms. By retaining the original gold labels, this augmented test set enables precise measurement of model performance on each error taxonomy, revealing whether debiasing reduces reliance on superficial patterns.

3.4 Evaluation Criteria

Performance is assessed across three complementary dimensions:

- Overall Inference Analysis: Measures general robustness on the augmented evaluation set.
- Per-Category Accuracy: Tracks improvements or regressions within individual error taxonomies.
- Qualitative Analysis: Manual inspection of misclassifications highlights subtle weaknesses not captured by accuracy scores.

Chapter 4

Experimentation

The experimental evaluation applies the methodological framework described in the previous chapter to the BERT model (Devlin et al., 2018). The aim is to understand the degree of improvement in NLI tasks when bias in the dataset is removed.

4.1 Experiment Setup

4.1.1 Models

Experiments are conducted by fine tuning Bert-base-uncased LLM model(Devlin et al., 2018). This version of the model has not been exposed to MNLI during pre-training, making it a fair backbone for testing debiasing effects. BERT is a standard choice for NLI due to its strong performance, bidirectional attention, and manageable computational cost. The BERT model is fine tuned on two datasets creating two versions which will be compared. The hyperparameters and model configurations used for fine tuning can be found in table 4.1.

4.1.2 Dataset

The datasets used for fine tuning are :-

- MNLI Dataset(Williams et al., 2018) - A large-scale benchmark for natural language inference containing premise–hypothesis pairs annotated with entailment, neutral, or contradiction. This dataset contains inherent biases which the models learn using shortcuts. The first

Hyperparameter	**Value**
Model Configuration	
Model ID	`bert-base-uncased`
Number of labels	3
Hidden dropout probability	0.2
Attention dropout probability	0.1
Training Arguments	
Number of epochs	5
Train batch size	32
Eval batch size	16
Learning rate	1e-5
Max sequence length	256
Weight decay	0.01
Warmup steps	2000
Logging steps	200
LR scheduler	`constant_with_warmup`
Early stopping patience	2
FP16	True

Table 4.1: Hyperparameters used for model training and evaluation.

row of table 4.2 presents two illustrative cases. In the first case, the premise and hypothesis exhibit high lexical overlap and the gold label is Entailment; such examples encourage the model to learn the spurious rule that high overlap implies entailment. In the second, the gold label is Contradiction and the hypothesis contains the word "not," which can lead the model to adopt the shortcut that the presence of negation signals contradiction.

- Debiased MNLI Dataset (Wu et al., 2022) - a version of MNLI augmented with synthetic premise–label–hypothesis triples produced by a trained generator (GPT-2) and then filtered (consistency filtering + z-filtering) to remove examples that contribute to known spurious correlations (e.g., hypothesis-only, lexical-overlap, contradiction-word biases). A few representative examples from the dataset are shown in the second row of Table 4.2. In the first two cases, although the lexical overlap between premise and hypothesis is very high, the gold label is not Entailment, preventing models from relying on lexical overlap as a shortcut. In the next two cases, despite the presence of the word "not," the gold label is not Contradiction, discouraging models from adopting the spurious rule that negation always signals contradiction.

The models trained on the above datasets are evaluated on the below datasets.

- Negation Prefixed MNLI Test set (Sullivan, 2024) - A version of MNLI dataset augmented by affixing a negation prefix to the hypothesis thereby flipping the label.
- Syntactic Augmented Dataset (Min et al., 2020) - a dataset created to challenge-style NLI pairs by applying transformations like subject–object inversion and passivization to sentences from MNLI in order to test and improve model sensitivity to syntactic structure.
- Synonym-augmented MNLI (Nguyen, 2023) - a transformed MNLI test set created by replacing words in hypotheses with synonyms to test lexical robustness.

Dataset	**Premise**	**Hypothesis**	**Label**
Original	Perhaps he should clear out even farther and head for California.	Maybe he should head for California to clear out even farther?	Entailment
	The herbal extract shows success in treating mild dementia and preventing Alzheimer's memory loss.	The herbal extract is not very good at treating dementia.	Contradiction
Debiased	I do the SERVICE, monsieur.	Only I do the service.	Neutral
	Like the sculptures on the temples, the eloquence of dance is a means of transmitting the messages of the holy scriptures and adventures of the great Hindu epics to its listeners.	Unlike the sculptures on the temples, the eloquence of dance is not a means of transmitting the adventures of the great Hindu epics.	Contradiction
	What is not apparent is that I have a bladder problem.	It's not apparent that I have a bladder problem.	Entailment
	General consensus of those providing input said that the great Hindu epics are meaningful to its listeners.	People who did not provide input said that the great Hindu epics are meaningful to its listeners.	Neutral

Table 4.2: Examples of premise–hypothesis pairs from the Original and Debiased datasets.

- HANS Dataset (McCoy et al., 2019) - a stress-test dataset that pairs hypotheses with premises engineered to break common lexical heuristics (e.g. lexical overlap, subsequence, constituent) while preserving entailment, thus revealing when models exploit superficial cues rather than true inference.

Refer to table A.1 in the appendix for examples from each dataset.

4.2 Results and Observations

Dataset	Model	Accuracy	F1 macro
HANS	base	62.40%	56.86%
HANS	debiased	**64.39%**	**59.61%**
Negation Prefixed MNLI Test set	base	**49.96%**	**44.62%**
Negation Prefixed MNLI Test set	debiased	48.84%	43.36%
Passivized Augmented Set	base	**93.66%**	**93.55%**
Passivized Augmented Set	debiased	90.12%	90.01%
Inverted Augmented Set	base	3.87%	2.48%
Inverted Augmented Set	debiased	**9.14%**	**5.58%**
Synonym-augmented MNLI	base	66.83%	70.24%
Synonym-augmented MNLI	debiased	**68.10%**	**70.70%**

Table 4.3: Comparison of baseline and debiased models across different evaluation datasets. Bold values indicate the better-performing model for each metric.

The experiments demonstrate both the promise and the limitations of dataset debiasing for NLI. Across all evaluations, a consistent pattern emerged: debiasing improved robustness to well-known lexical heuristics, yet sometimes reduced performance on constructions where those heuristics coincided with the gold label. On average, the model trained on the debiased dataset achieved higher accuracy than the baseline trained on standard MNLI. Importantly, these gains appear to stem from a deeper reliance on semantic understanding of the premise–hypothesis pair rather than on superficial cues such as lexical overlap. At the same time, the debiased model exhibited lower prediction confidence overall: while the baseline consistently produced near-maximal probabilities (0.997), the debiased model's outputs were more moderate (0.933–0.952). This suggests that debiasing not only reduces shortcut learning but also yields models that are more conservative and less overconfident in their predictions.

4.2.1 Category wise analysis

Negation Prefixed MNLI Test set

A more nuanced picture emerged on the negation stress tests. As shown in Table 4.3, the debiased model achieved slightly lower accuracy and macro-F1 compared to the MNLI baseline. This can be attributed to the baseline model's reliance on negation-related heuristics present in the original MNLI data, which artificially boosted its performance on such examples. In contrast, the debiased model, trained on data where these artifacts were reduced, no longer benefited from such shortcuts, leading to a modest drop in overall scores. However, analysis of the confusion matrices revealed that the debiased model was less prone to misclassifying negated hypotheses as Contradiction, instead distributing probability more appropriately toward the Neutral class. This suggests genuine progress in mitigating a well-documented bias, even if it comes at the cost of reduced balance across classes.

Example

> **Premise:** "If that investor were willing to pay extra for the security of limited downside, she could buy put options with a strike price of $98, which would lock in her profit on the shares at $18, less whatever the options cost."
>
> **Hypothesis:** "It is not true that the strike price could be $8."
>
> **Gold label:** Entailment
> **Base prediction:** Contradiction (prob $\approx$ 0.9999999)
> **Debiased prediction:** Neutral (prob $\approx$ 0.763)
>
> **Interpretation:** The baseline again treats the negated hypothesis as a contradiction. The debiased model resists this heuristic, shifting probability mass toward Neutral. Although it does not reach the gold label, the change demonstrates reduced systematic bias in the presence of negation.

Syntactic Augmented Dataset

The syntactic transformation tests (passivization and inversion) revealed contrasting outcomes. As shown in Table 4.3, the debiased model achieved higher accuracy and macro-level metrics on the inverted augmented dataset, although absolute scores for both models remained low. In this setting, the debiased model recovered substantially more examples than the baseline (78 vs. 14 debiased-only correct), while the majority of examples (1090/1215) were still misclassified by both models. Detailed analysis showed that many inverted examples retained high lexical overlap between premise and hypothesis (mean overlap $\approx$0.65), which misled the baseline into treating subject–object reversals as entailments. By reducing reliance on lexical overlap, the debiased model was better able to resist this heuristic and demonstrated improvements concentrated in these high-overlap cases, a detailed example of such a case is given below. This result aligns with the intended effect of debiasing: reducing shortcut learning and improving sensitivity to syntactic roles and word order.

Example

> **Premise:** “A self-righteous tone suffuses this ... The author clearly feels he has exposed the dirty secrets that inspired his protagonist’s crypto agenda of reform.”
>
> **Hypothesis:** “His protagonist’s secrets exposed the author.”
>
> **Gold label:** Neutral
> **Base prediction:** Entailment
> **Debiased prediction:** Neutral
> **Lexical overlap:** 1.0 (all hypothesis tokens overlap with premise tokens).
>
> **Interpretation:** The baseline falls for lexical overlap and predicts Entailment. The debiased model avoids this shortcut and correctly predicts Neutral, supporting the claim that debiasing improves robustness to inversion, where overlap is misleading.

In contrast, for passivization, the baseline model outperformed the debiased model, particularly on the passivized split. Both models were correct on the majority of examples (1067/1215), but in cases where they disagreed, the baseline was more often correct (71 vs. 28 base-only correct). Here, the heuristic of mapping active to passive sentences as entailments was generally valid, and debiasing shifted the model toward more conservative predictions, frequently assigning Neutral instead of Entailment or Contradiction. As a result, the debiased model lost accuracy on this subset, illustrating that debiasing can sometimes suppress heuristics that are aligned with the ground truth, a detailed example can be found below.

Example

> **Premise:** “Get real!”
>
> **Hypothesis:** “Something perfectly normal was said by the person.” (passive-style rewrite)
>
> **Gold label:** Contradiction
> **Base prediction:** Contradiction
> **Debiased prediction:** Neutral
> **Lexical overlap:** 0.0
>
> **Interpretation:** The baseline captures the incompatibility and predicts Contradiction. The debiased model is more conservative, defaulting to Neutral, and loses accuracy. This illustrates a boundary shift, where debiasing removes a heuristic that in this case aligned with the gold label.

Synonym-augmented MNLI

On the Synonym-augmented dataset, where words in the hypothesis were replaced by near-synonyms, the debiased model consistently outperformed the baseline. Accuracy increased from 66.8% to 68.1% and macro-F1 from 66.4% to 67.9%. Lexical-overlap analysis indicated that these improvements were most pronounced in high-overlap cases, suggesting that the baseline relied heavily on token-level overlap when classifying. Confusion matrix inspection further revealed that the debiased model corrected many instances where the baseline underpredicted Entailment, successfully recovering examples that depended on semantic equivalence rather than exact lexical matches. Taken together, these results show that debiasing enhances robustness to lexical variation by reducing dependence on surface-level overlap and better capturing underlying meaning.

HANS Dataset

The HANS benchmark is specifically designed to probe models for reliance on superficial syntactic heuristics, such as lexical overlap, subsequence, and constituent matching. Overall, the debiased model achieved a clear improvement in both accuracy and macro-F1 on HANS, indicating reduced dependence on these shortcuts. In particular, the largest gains were observed on the lexical overlap subset, where the baseline frequently mislabeled examples with high word overlap as Entailment. By contrast, the debiased model showed greater sensitivity to underlying syntactic roles, correctly identifying non-entailments even when lexical overlap was high. This suggests that debiasing effectively helps in mitigating one of the most well-documented heuristics in NLI and encourages more structurally grounded reasoning.

Chapter 5

Conclusions

This study evaluated whether dataset-level debiasing enhances the robustness of Natural Language Inference (NLI) models when probed across established error taxonomies. The experiments demonstrated that debiasing consistently reduced reliance on superficial lexical overlap and improved performance on syntactic inversion, but at the same time led to declines in settings such as passivization where heuristics aligned with gold labels. These mixed outcomes highlight that biases in NLI datasets function as a double-edged sword: while they can create brittleness and shortcut learning, they may also encode correlations that are beneficial for certain constructions.

A central contribution of this work is to show that aggregate accuracy alone is insufficient to evaluate debiasing. Through systematic, category-wise probing of negation, synonym replacement, syntactic transformations, and adversarial stress tests, the study demonstrates that fine-grained diagnostics are essential to reveal where debiasing genuinely improves reasoning and where it suppresses useful cues.

The results suggest that future work requires more adaptive approaches to debiasing that selectively attenuate harmful heuristics while preserving those that support valid inference. This refinement of strategies will enable models to leverage linguistic regularities in ways that align with true semantic reasoning.

References

Belinkov, Y., Gehrmann, S., Pavlick, E., Rush, A., Sachan, M. and Poole, D. (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', Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding'.

Glockner, M., Shwartz, V. and Goldberg, Y. (2018), Breaking NLI systems with sentences that require simple lexical inferences, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R. and Smith, N. A. (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', Association for Computational Linguistics.

He, H., Zha, S. and Wang, H. (2019), Unlearn dataset bias in natural language inference by fitting the residual, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics.

McCoy, R. T., Pavlick, E. and Linzen, T. (2019), 'Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference'.

Min, J., McCoy, R. T., Das, D., Pitler, E. and Linzen, T. (2020), 'Syntactic data augmentation increases robustness to inference heuristics'.

Naik, A., Ravichander, A., Sadeh, N., Rose, C. and Neubig, G. (2018), Stress test evaluation for natural language inference, *in* 'Proceedings of the 27th International Conference on Computational Linguistics', Association for Computational Linguistics.

Nguyen, D. T. (2023), 'Dissecting vocabulary biases datasets through statistical testing and automated data augmentation for artifact mitigation in natural language inference'.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R. and Van Durme, B. (2018), Hypothesis only baselines in natural language inference, *in* 'Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics', Association for Computational Linguistics.

Sullivan, M. (2024), It is not true that transformers are inductive learners: Probing NLI models with external negation, *in* Y. Graham and M. Purver, eds, 'Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, St. Julian's, Malta, pp. 1924–1945.
URL: *https://aclanthology.org/2024.eacl-long.116/*

Utama, P. A., Moosavi, N. S. and Gurevych, I. (2020), Towards debiasing nlu models from unknown biases, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics.

Williams, A., Nangia, N. and Bowman, S. (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)', Association for Computational Linguistics, pp. 1112–1122.
URL: *http://aclweb.org/anthology/N18-1101*

Wu, Y., Gardner, M., Stenetorp, P. and Dasigi, P. (2022), 'Generating data to mitigate spurious correlations in natural language inference datasets'.

Appendix A

Category wise examples

Error Taxonomy	Premise	Hypothesis	Label
Inversion	The carriage creaked and groaned in protest; the whole thing wavering with stress.	A lot of noise made the carriage.	Neutral
Passivization	For the world, a few changes would be needed.	No changes were needed by the world.	Contradiction
Synonym	The athlete won the race easily.	The athlete triumphed in the race easily.	Entailment
Lexical-Overlap HANS	Every student read the chapters assigned to them carefully.	Every student read the chapters carefully.	Non-entailment
Negation Prefixed	Sorry but that's how it is.	It is not true that this is how things are and there are no apologies about it.	Entailment

Table A.1: Illustrative augmented examples under different error taxonomies.

Source Code

Source code to the experiment can be found at - Link to Github