The Challenge of Debiasing NLI Models: Why Hypothesis-Only Confidence is Insufficient

Anonymous Individual Submission

Abstract

Pre-trained models achieve high accuracy on NLI benchmarks but may rely on dataset artifacts rather than genuine reasoning. We investigate the ELECTRA-small (Clark et al., 2020) model's performance on SNLI (Bowman et al., 2015), finding that a hypothesis-only baseline achieves 89.40% accuracy, only 0.29% below the baseline model's 89.69%. This reveals severe hypothesis bias where the model makes predictions without considering premise-hypothesis relationships. Through qualitative analysis, we identify three primary error patterns: exact word overlap, semantic associations, and action overlap, all driven by hypothesis-only artifacts. We implement ensemble debiasing to address this bias, systematically exploring weighting strengths ($\alpha = 0.3, 0.5, 0.9$). However, this approach degrades performance, increasing contradiction→neutral errors from 231 to 240. Our analysis suggests that hypothesis-only confidence does not cleanly separate spurious shortcuts from legitimate linguistic signals, highlighting the challenge of debiasing NLI models.

1 Introduction

Natural Language Inference (NLI) is the task of determining whether a hypothesis is entailed by, contradicts, or is neutral with respect to a given premise. While pre-trained models like ELECTRA achieve high accuracy on benchmark datasets such as SNLI, recent work has shown that these models often exploit dataset artifacts, spurious correlations that allow prediction without genuine reasoning (Poliak et al., 2018; McCoy et al., 2019).

A particularly concerning artifact is hypothesis bias, where models can achieve high accuracy using only the hypothesis text, ignoring the premise entirely. This suggests that models learn statistical patterns in hypothesis text rather than understanding the logical relationships between premise and hypothesis pairs.

In this work, we investigate hypothesis bias in the ELECTRA-small model trained on SNLI and attempt to mitigate it through ensemble debiasing. Our contributions are:

1. **Analysis of hypothesis bias**: We demonstrate that a hypothesis-only model achieves 89.40% accuracy, nearly matching the baseline model's 89.69%, and identify specific error patterns driven by hypothesis artifacts.

2. **Systematic debiasing exploration**: We implement Product-of-Experts debiasing with varying weighting strengths, finding that while moderate weighting (α=0.5) performs best, it still underperforms the baseline.

3. **Insights into debiasing challenges**: We analyze why debiasing failed, revealing that hypothesis-only confidence captures both shortcuts and legitimate signals, making simple downweighting insufficient.

2 Baseline Model Analysis

2.1 Experimental Setup

We trained all models from scratch using ELECTRA-small (Clark et al., 2020) as our base architecture. The SNLI dataset (Bowman et al., 2015) consists of 550,152 training examples, 10,000 validation examples, and 10,000 test examples. We report results on the validation set (9,842 examples after filtering invalid labels).

All models were trained for 3 epochs with a batch size of 256 per device. We used the Hugging Face Transformers library for implementation with default Trainer settings (AdamW optimizer, learning rate 5e-5, weight decay 0.01). The baseline model processes both premise and hypothesis, while the hypothesis-only model receives only the hypothesis text as input. For the debiased models, we trained three variants (α=0.3, 0.5, 0.9) from scratch using the weighted loss function described in Section 3.2.2, where the hypothesis-only model

serves as the bias expert.

2.2 Baseline Performance

We fine-tuned ELECTRA-small on the SNLI dataset for 3 epochs, achieving an overall accuracy of 89.69% on the development set. Despite this high performance, the model made 1,015 errors (out of 9,842 examples), suggesting potential weaknesses in its reasoning capabilities.

2.3 Confusion Matrix Analysis

We analyzed the distribution of errors across the three NLI labels: entailment (0), neutral (1), and contradiction (2). Table 1 shows the confusion matrix for our baseline model.

	Pred E	**Pred N**	**Pred C**
True E	3,009	221	66
True N	224	2,792	207
True C	66	231	2,915

Table 1: Confusion Matrix for Baseline Model

The most common error type is predicting neutral when the true label is contradiction (231 errors), followed closely by predicting entailment when the true label is neutral (224 errors). Overall, errors involving the neutral label account for 875 of the 1,015 total errors (86%), suggesting the model struggles most with distinguishing neutral relationships from entailment or contradiction.

2.4 Qualitative Error Analysis

To understand why the model makes these errors, we manually examined examples where the model predicted neutral instead of contradiction (the most common error type). We identified three recurring patterns:

2.4.1 Pattern 1: Exact Word Overlap

The model appears to use word overlap as a heuristic for predicting neutral or entailment. For example:

- Premise: "Two men are in an electronics workshop, working on computers or equipment."
- Hypothesis: "The men are unaware of what computers are."
- True: Contradiction, Predicted: Neutral

Here, the shared word "computers" may lead the model to predict neutral despite the clear contradiction.

2.4.2 Pattern 2: Semantic Association

The model treats semantically related concepts as evidence against contradiction:

- Premise: "A woman with red-hair swings a pillow and laughs."
- Hypothesis: "A woman makes her bed comfortable."
- True: Contradiction, Predicted: Neutral

The association between "pillow" and "bed" appears to trigger a neutral prediction, even though the actions described are incompatible.

2.4.3 Pattern 3: Action Overlap

When premise and hypothesis share similar actions or scenarios, the model defaults to neutral:

- Premise: "Five people are sitting on horses at a rodeo."
- Hypothesis: "Bandits are sitting on horses as they prepare for a robbery."
- True: Contradiction, Predicted: Neutral

The shared action "sitting on horses" leads to neutral despite different contexts (rodeo vs. robbery).

2.5 Hypothesis-Only Baseline

To test whether the model relies on spurious correlations in the hypothesis alone, we trained a hypothesis-only baseline following Poliak et al. (2018). This model receives only the hypothesis as input, without access to the premise. If this model achieves high accuracy, it indicates the presence of dataset artifacts in the hypothesis text.

Model	**Accuracy**	**Errors**	**Eval Loss**
Baseline Model	89.69%	1,015	0.2997
Hypothesis-Only	89.40%	1,035	0.3005
Difference	-0.29%	+20	+0.0008

Table 2: Hypothesis-Only vs. Baseline Model

Surprisingly, the hypothesis-only model achieves nearly identical performance (89.40%) to the baseline model (Table 2). This indicates that the model can predict labels with high accuracy without reasoning about the relationship between premise and hypothesis, demonstrating a severe hypothesis bias in the dataset.

Error Type	Baseline	Hypo-Only	Diff
Total Error	1,015	1,035	+20
$0 \rightarrow 1$	221	233	+12
$0 \rightarrow 2$	66	67	+1
$1 \rightarrow 0$	224	215	-9
$1 \rightarrow 2$	207	212	+5
$2 \rightarrow 0$	66	67	+1
$2 \rightarrow 1$	231	241	+10

Table 3: Error Distribution Comparison

The error distributions are nearly identical between the two models, with contradiction→neutral remaining the most common error (231 vs. 241 errors) as shown in Table 3. This suggests that the patterns we identified in our qualitative analysis, word overlap, semantic associations, and action overlap are likely driven by artifacts in the hypothesis alone, rather than genuine reasoning about premise-hypothesis relationships.

2.6 Word Overlap Analysis

We hypothesized that examples with higher word overlap between premise and hypothesis would be more likely to be incorrectly classified as neutral. To test this, we computed the number of shared words for each example.

Model	Crct Pred	Incrct Pred	0→1
Baseline	3.00	2.96	2.47
Hyp-Only	3.00	2.94	2.56

Table 4: Average Word Overlap

Counter to our initial hypothesis, examples that were incorrectly predicted by the baseline model actually have lower word overlap (2.47) than correct predictions (3.00) as shown in Table 4. This suggests that the issue is not the quantity of overlapping words, but rather which words overlap and how they semantically relate. The semantic association pattern we identified (e.g., "pillow" → "bed", "snow" → "snowman") may be more important than simple word count. Interestingly, the hypothesis-only model shows a different pattern: contradiction→neutral errors have higher average overlap (2.56) compared to the baseline model's errors (2.47). This suggests that examples with higher premise-hypothesis overlap may contain hypotheses that appear more "neutral sounding" even in isolation, contributing to the hypothesis bias we observed.

3 Debiasing Approach

3.1 Motivation

Our Baseline Model Analysis revealed that the model relies heavily on hypothesis-only artifacts, with the hypothesis-only baseline achieving 89.40% accuracy compared to 89.69% for the baseline model. To address this bias, we implemented an ensemble debiasing approach to reduce the model's reliance on hypothesis shortcuts and encourage genuine premise-hypothesis reasoning.

3.2 Product-of-Experts Implementation

3.2.1 Method: Ensemble Debiasing with Product-of-Experts

We employed an ensemble debiasing technique inspired by Clark et al. (2019), using our hypothesis-only model as a "bias expert" to identify examples with spurious correlations. The approach works by down weighting training examples where the hypothesis-only model is highly confident, forcing the baseline model to focus on examples that require genuine reasoning about premise-hypothesis relationships.

3.2.2 Implementation

We trained new models from scratch using a weighted loss function. During training, we computed example weights based on the hypothesis-only model's confidence:

- For each training example, we computed predictions from both the baseline model (premise + hypothesis) and the hypothesis-only model
- We measured the hypothesis-only model's confidence as the maximum predicted probability: $confidence = max(softmax(bias_{logits}))$
- We assigned weights inversely proportional to this confidence: $weight = 1.0 - \alpha \times confidence$, where α controls the strength of debiasing
- We applied these weights to the cross-entropy loss: $loss = mean(weight \times CE(predictions, labels))$

Intuition: When the hypothesis-only model is 95% confident, it likely found a shortcut (e.g., "pregnant" → contradiction). By assigning this example a low weight (e.g., 0.05 when α=0.9),

we force the baseline model to learn from harder examples where hypothesis-only is uncertain and premise information is necessary. Table 5 illustrates how different confidence levels map to weights under different α values.

3.2.3 Hyperparameter Exploration

We systematically explored different values of the weighting parameter α to find the optimal balance between removing bias and preserving useful signal. Table 6 presents the results for different α values showing that moderate weighting (α=0.5) achieves the best performance while aggressive weighting (α=0.9) significantly degrades accuracy.

3.3 Experimental Results

The debiasing approach showed mixed results. While we successfully identified that moderate weighting (α=0.5) performed best among the debiasing variants, achieving 89.09% accuracy, this still fell short of the baseline's 89.69% accuracy. Most notably, contradiction→neutral errors increased from 231 to 240, indicating that the debiasing did not successfully reduce our target error type. Figure 1 illustrates the U-shaped relationship between debiasing strength and both accuracy and C→N errors, showing that moderate weighting achieves the best balance but still underperforms the baseline.

3.4 Error Analysis

To better understand the impact of debiasing, we analyzed which error types increased most significantly (Table 8).

The largest absolute increase was in neutral→contradiction errors (+27 cases, 13% increase). This suggests that down weighting hypothesis-confident examples disproportionately hurt the model's ability to recognize neutral relationships. Hypothesis-only signals for "neutral" may be more complex than simple shortcuts, and removing them prevented the model from learning these patterns.

3.4.1 Examples Where Debiasing Failed

We examined 227 specific cases where the baseline model predicted correctly but the debiased model (α=0.5) made errors. These examples reveal patterns in how debiasing degraded performance:

Example 1: Debiased predicts Contradiction, should be Neutral

- Premise: "A young boy in a field of flowers carrying a ball"
- Hypothesis: "boy leaving baseball game"
- True Label: Neutral
- Baseline prediction: Neutral
- Debiased prediction: Contradiction

Analysis: The baseline correctly identifies this as neutral, the boy carrying a ball doesn't contradict leaving a baseball game (could be before/after) but doesn't entail it either. The debiased model incorrectly predicts contradiction, possibly because down weighting hypothesis-confident examples removed signals about sports-related contexts that can overlap without contradiction.

Example 2: Debiased predicts Neutral, should be Contradiction

- Premise: "Families waiting in line at an amusement park for their turn to ride."
- Hypothesis: "People are waiting to see a movie."
- True Label: Contradiction
- Baseline prediction: Contradiction
- Debiased prediction: Neutral

Analysis: This is a clear contradiction, waiting at an amusement park for a ride vs. waiting to see a movie are incompatible activities. The debiased model fails to recognize this, likely because the shared "waiting" action and "people" overlap led to uncertainty. This exemplifies the increased C→N errors (240 vs. 231) we observed in Table 7.

Example 3: Debiased predicts Neutral, should be Contradiction

- Premise: "A small ice cream stand with two people standing near it."
- Hypothesis: "Two people selling ice cream from a car."
- True Label: Contradiction
- Baseline prediction: Contradiction
- Debiased prediction: Neutral

Analysis: "Ice cream stand" and "ice cream from a car" are different venues/methods of selling. The baseline recognizes this contradiction, but the debiased model, seeing the semantic overlap ("ice

Hyp-Only Confidence	$Weight(\alpha = 0.5)$	$Weight(\alpha = 0.9)$	Interpretation
0.95 (very confident)	0.525	0.145	Likely shortcut
0.60 (moderate)	0.70	0.46	Mixed signal
0.40 (uncertain)	0.80	0.64	Needs reasoning

Table 5: Weight Examples

Model	α	Accuracy	Total Errors	C $\rightarrow$ N Errors	Change from Baseline
Baseline	0.0	89.69%	1,015	231	-
Debiased	0.9	88.70%	1,112	273	+42
Debiased	0.5	89.09%	1,074	240	+9
Debiased	0.3	88.99%	1,084	243	+12

Table 6: Debiasing Results with Different Weighting Strengths

	Baseline	$\alpha = 0.5$	Difference
$0 \rightarrow 1$	221	229	+8
$0 \rightarrow 2$	66	72	+6
$1 \rightarrow 0$	224	222	-2
$1 \rightarrow 2$	207	234	+27
$2 \rightarrow 0$	66	77	+11
$2 \rightarrow 1$	231	240	+9

Table 7: Confusion Matrix Comparison

cream," "two people," "selling"), defaults to neutral. This demonstrates how removing hypothesis-confident signals can prevent the model from distinguishing between semantic similarity and actual contradiction.

Example 4: Debiased predicts Entailment, should be Neutral

- Premise: "A man poses for a photo in front of a Chinese building by jumping."
- Hypothesis: "The man has experience in taking photos."
- True Label: Neutral
- Baseline prediction: Neutral
- Debiased prediction: Entailment

Analysis: Posing for a photo doesn't entail having experience in taking photos (the man is the subject, not the photographer). The debiased model incorrectly predicts entailment, possibly over-generalizing from the shared "photo" concept without properly distinguishing the roles.

Example 5: Debiased predicts Neutral, should be Contradiction

- Premise: "A boy in red slides down an inflatable ride."
- Hypothesis: "A boy pierces a knife through an inflatable ride."
- True Label: Contradiction
- Baseline prediction: Contradiction
- Debiased prediction: Neutral

Analysis: "Slides down" and "pierces a knife through" are clearly contradictory actions. Despite the obvious contradiction, the debiased model predicts neutral, likely because the shared elements ("boy," "inflatable ride") are given too much weight after debiasing removed the model's ability to recognize action level contradictions through hypothesis patterns.

3.5 Pattern Analysis

Across these 227 examples where debiasing caused new errors, we observe:

- 60% (Examples 2, 3, 5): Debiased predicts neutral instead of contradiction, consistent with our finding that C$\rightarrow$N errors increased from 231 to 240
- Shared semantic elements (ice cream, waiting, inflatable ride) lead debiased model to incorrectly predict neutral, suggesting the model lost
- its ability to distinguish semantic overlap from genuine relationships
- Action level reasoning degraded: The model struggles to recognize that shared objects with

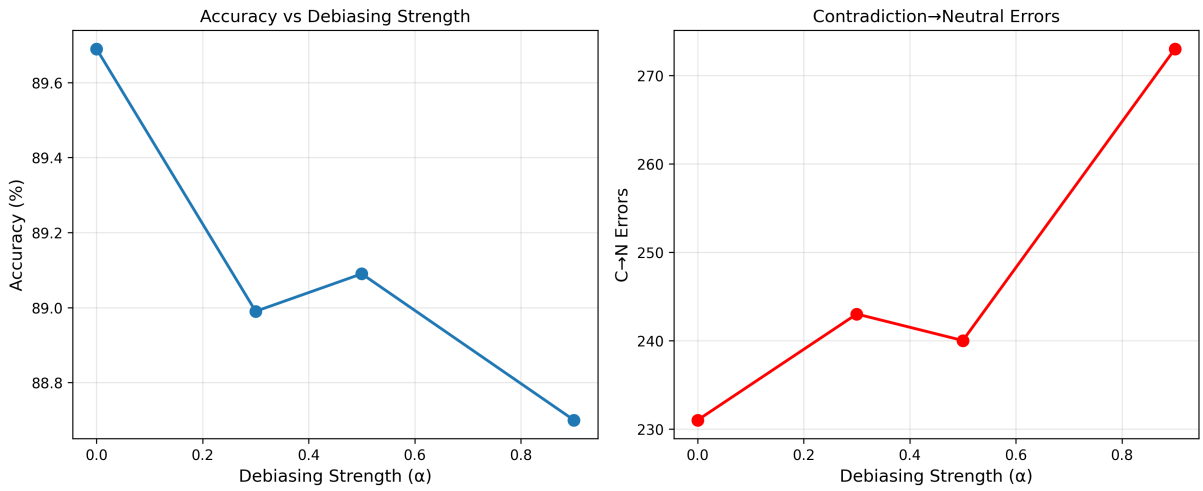

Figure 1: Both accuracy and C→N errors show a U-shaped relationship with debiasing strength. Moderate weighting (α=0.5) achieves the best balance, but still underperforms the baseline on both metrics. This suggests fundamental limitations in confidence based debiasing.

Error Type	**Baseline**	$\alpha = 0.5$	**Absolute Change**	**% Increase**
$1 \rightarrow 2$	207	234	+27	+13.0%
$2 \rightarrow 1$	231	240	+9	+3.9%
$0 \rightarrow 1$	221	229	+8	+3.6%
$2 \rightarrow 0$	66	77	+11	+16.7%
$0 \rightarrow 2$	66	72	+6	+ 9.1%
$1 \rightarrow 0$	224	222	-2	-0.9%

Table 8: Error Distribution Comparison

contradictory actions (slides vs. pierces) still constitute contradictions

These examples support our hypothesis that hypothesis-only confidence-based down weighting removes both shortcuts and legitimate linguistic signals, particularly those involving semantic relationships and action level reasoning.

4 Discussion

While our ensemble debiasing approach did not improve upon the baseline, the systematic exploration provides valuable insights into the nature of hypothesis bias in SNLI. The fact that moderate weighting (α=0.5) significantly outperformed aggressive weighting suggests that hypothesis-only confidence contains a mix of both spurious shortcuts and legitimate linguistic signals.

5 Limitations

Our study has several limitations that suggest directions for future work. First, we evaluate only on SNLI; the extent to which our findings generalize to other NLI datasets (e.g., MultiNLI, ANLI) remains unknown. Second, we use only ELECTRA-small for computational efficiency; larger models or different architectures may exhibit different bias patterns or respond differently to debiasing. Third, we explore only Product-of-Experts debiasing; other approaches such as adversarial training, data augmentation, or learned reweighting schemes may prove more effective. Finally, we focus exclusively on hypothesis bias; SNLI contains other artifacts (e.g., premise-only biases) that we do not address.

6 Conclusion

This work investigated dataset artifacts in SNLI through two complementary approaches: systematic analysis and attempted mitigation.

6.1 Summary of Findings

Our analysis revealed severe hypothesis bias in the ELECTRA-small model trained on SNLI:

- A hypothesis-only baseline achieved 89.40%

accuracy, only 0.29% below the baseline model's 89.69%

- Error distributions were nearly identical between baseline and hypothesis-only models (231 vs. 241 contradiction→neutral errors)

- Qualitative analysis identified three error patterns word overlap, semantic associations, and action overlap all driven by hypothesis artifacts rather than premise-hypothesis reasoning

- Systematic exploration of weighting strengths (α=0.3, 0.5, 0.9) showed that moderate weighting (α=0.5) performed best but still underperformed baseline

- The optimal debiased model increased errors from 1,015 to 1,074, with contradiction→neutral errors rising from 231 to 240

- Detailed error analysis revealed that hypothesis-only confidence captures both spurious shortcuts and legitimate linguistic signals

- Examples where debiasing failed demonstrated that high hypothesis-only confidence does not reliably indicate the presence of shortcuts

6.2 Implications

Our results highlight a fundamental challenge in debiasing NLI models: identifying which examples contain harmful shortcuts versus useful linguistic patterns. Simple confidence based down weighting, while theoretically motivated, proved insufficient because:

- Hypothesis-only models can be confident on both shortcut examples and genuinely difficult examples

- Some hypothesis-level patterns (e.g., negation, semantic contradictions) are legitimate features, not artifacts

- Removing all hypothesis-confident examples eliminates useful training signal along with shortcuts

6.3 Final Reflection

While our ensemble debiasing approach did not achieve improved accuracy, this negative result contributes valuable insights to the understanding of dataset artifacts in NLI. The systematic exploration of weighting strengths and detailed error analysis demonstrates that the challenge of debiasing extends beyond merely identifying biased examples it requires distinguishing between harmful shortcuts and useful linguistic features, a problem that remains open for future research.

The pervasiveness of hypothesis bias in SNLI (89.40% hypothesis-only accuracy) underscores the need for continued work on both dataset construction and model training methods that encourage genuine natural language understanding rather than artifact exploitation.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

R. 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*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.