

The Architecture Was Doing the Work: Why NLI Debiasing Results Do Not Generalise*

Prabhjot Singh

The University of Texas at Austin

prabhjot.singh@utexas.edu

Abstract

NLI debiasing research evaluates interventions on single architectures and datasets, leaving a critical assumption untested: do improvements reflect method quality or model capacity? We run 44 controlled experiments: eleven debiasing strategies across two architecturally distinct models (ELECTRA-small, RoBERTa-base) and two corpora (SNLI, MNLI), under identical conditions with Wilson 95% confidence intervals throughout. On ELECTRA-small, which achieves 87.4% in-domain accuracy, debiasing fails completely: all ten debiasing interventions fall within the baseline confidence interval on HANS (CoV=0.75%). On RoBERTa-base, methods show genuine sensitivity (CoV=10.76%, approximately $14\times$ higher), yet the median intervention trends toward degrading OOD accuracy on SNLI, short of conventional significance (Wilcoxon $p=0.065$), and method rankings across corpora are anti-correlated for ELECTRA (Spearman $\rho=-0.42$) but stable for RoBERTa ($\rho=0.75$). Contrastive augmentation swings 5.6 pp across corpora (+3.59 pp MNLI, -2.02 pp SNLI). The positive robustness-accuracy correlation on RoBERTa ($r=0.848$) inverts the assumed tradeoff: methods that improve HANS also preserve in-domain accuracy, while aggressive debiasers collapse both. The ELECTRA failure is not a pipeline artefact; the same codebase produces real gains on RoBERTa. It is a capacity failure: debiasing presupposes representational capacity that small architectures lack, a precondition the literature neither states nor tests.

1 Introduction

Every published NLI debiasing method carries an assumption nobody has tested: that a technique which works on one model and one dataset will work in general. We ran the experiment the field

has skipped and the answer is no. Applying the same eleven debiasing strategies, under identical conditions, to two architecturally distinct models, we find that method effectiveness is determined almost entirely by which architecture receives it, not by the method itself. The same code, seed, and data produce real improvements on one model and complete statistical silence on the other.

This is possible because the NLI debiasing literature has an evaluation blindspot it has not noticed. Methods are validated on single architectures, typically RoBERTa-large or BERT-base, on a single training corpus, without confidence intervals, paired significance tests, or cross-architecture validation. The community accumulates apparent progress while measuring a conflation of method effect and architecture effect that no single-architecture study can disentangle.

McCoy et al. (2019) demonstrated that NLI models exploit lexical heuristics rather than learning genuine inference; subsequent work proposed ensemble debiasing (He et al., 2019), Product-of-Experts (Clark et al., 2019), Debaised Focal Loss (Karimi Mahabadi et al., 2020), and data augmentation (Nie et al., 2020) as correctives. Every one was validated predominantly on one architecture. None asked whether the correction requires a minimum level of model capacity to function.

We ask that question across 44 controlled experiments: eleven debiasing strategies applied to ELECTRA-small (Clark et al., 2020) and RoBERTa-base (Liu et al., 2019) on both SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), evaluated on HANS (McCoy et al., 2019). Every run shares an identical seed, optimiser, and evaluation protocol; every HANS accuracy carries a Wilson 95% CI; every cross-method comparison uses McNemar tests with Bonferroni correction. This is the first study to isolate the architecture variable across this many methods simultaneously.

On ELECTRA-small (87.4% in-domain on

*Code at: <https://github.com/prabhjotschugh/nli-debiasing-generalisation-study>

SNLI, not a trivially weak baseline), every intervention fails: CoV=0.75%, all ten methods within the baseline CI (Wilcoxon $p = 0.070$ SNLI, $p = 0.557$ MNLI). On RoBERTa-base, methods move but inconsistently: the median trends negative though not conventionally significant, and DFL ($\gamma = 3.0$) collapses in-domain accuracy by 22.96 pp while destroying robustness. Rankings reverse across corpora for ELECTRA but are stable for RoBERTa, underscoring that even RoBERTa’s genuine method sensitivity does not translate into architecture-independent rankings. The robustness-accuracy correlation on RoBERTa is strongly positive ($r = 0.848$), inverting the assumed tradeoff.

Our contributions: (1) the first controlled cross-architecture, cross-corpus evaluation of eleven NLI debiasing methods with full statistical reporting; (2) a novel failure mode where bias-model-dependent methods silently collapse to plain cross-entropy under insufficient model capacity; (3) dataset–method interactions (5.6 pp reversal) invisible to single-corpus evaluation; and (4) a positive robustness–accuracy correlation ($r = 0.848$) inverting the assumed tradeoff. These findings motivate four evaluation standards we argue should be non-negotiable for future debiasing work: cross-architecture validation on at least two structurally distinct models, Wilson confidence intervals with paired significance tests, model capacity reported as a moderating variable, and HANS per-heuristic breakdowns rather than a single aggregate score. We return to these in §4.

2 Experimental Setup

Models. ELECTRA-small

(google/electra-small-discriminator; 14M parameters) is pretrained via Replaced-Token-Detection, optimising for local token-level prediction.

RoBERTa-base (FacebookAI/roberta-base; 125M parameters) is pretrained via Masked Language Modelling, encouraging broader contextual representations. Both are fine-tuned as 3-class sequence classifiers.

Datasets. Training: SNLI (Bowman et al., 2015) (549K) and MNLI (Williams et al., 2018) (393K); unresolved labels filtered. In-domain evaluation uses SNLI test and MNLI validation-matched. OOD robustness uses HANS (McCoy et al., 2019): 30K binary examples across lexical overlap, subsequence, and constituent heuristics (10K each); per-heuristic breakdowns reported throughout.

Methods. 11 strategies share a unified codebase to eliminate confounders, grouped into 4 families:

- *Degenerate reference* - **Hypothesis-Only**: trained on hypothesis text alone; establishes the artifact ceiling.
- *Loss modification* - **Label Smoothing** ($\epsilon = 0.15$); **Confidence Penalty** ($\beta = 0.1$) (Pereyra et al., 2017) via $L = \text{CE} - \beta \cdot H(p)$, first applied here to NLI debiasing.
- *Bias-model-based* - **Ensemble Debiasing** ($\alpha = 0.5$) (He et al., 2019); **Product-of-Experts** ($\alpha = 0.5$) (Clark et al., 2019); **DFL** (Karimi Mahabadi et al., 2020) at aggressive ($\gamma = 3.0$, 1 ep) and conservative ($\gamma = 2.0$, 2 ep) settings, epoch counts fixed a priori to bound aggressive down-weighting rather than tuned per run. All use Hypothesis-Only as the bias model.
- *Data augmentation* - **ANLI Augmentation** (25% adversarial (Nie et al., 2020)); **Contrastive Augmentation** (150K rule-based contradictions via negation, antonym, quantifier, number swap); **Aggressive Augmentation** (same pipeline at 30% train size).

Training and statistics. All runs: AdamW, $\text{lr} = 2\text{e-}5$, weight decay 0.01, linear warmup (10%), seed 42 (Python/NumPy/PyTorch/CUDA), BF16 where supported. Default 3 epochs; DFL uses 1–2 epochs; contrastive augmentation uses 5 epochs. Batch sizes 32–256 per GPU memory; dual-input methods use halved batches. All HANS accuracies carry Wilson 95% CIs; Wilcoxon signed-rank tests assess systematic improvement (ten methods vs. baseline per model $\times$ corpus); Spearman ρ measures ranking consistency; McNemar tests with Bonferroni correction for pairwise significance. The *premise-irrelevance gap* (full-model minus Hypothesis-Only HANS accuracy) diagnoses premise conditioning capacity.

3 Results

3.1 Baseline Performance and Premise-Irrelevance Gap

Table 1 reports all 44 results. Both models are fully functional in-distribution: ELECTRA achieves 87.41% on SNLI ([86.74, 88.05]) and 79.37% on MNLI ([78.56, 80.16]); RoBERTa achieves 91.50% ([90.93, 92.04]) and 87.96% ([87.30, 88.59]). On HANS, ELECTRA collapses to near-random: 49.84% on SNLI and 51.87%

Method	ELECTRA-small				RoBERTa-base			
	SNLI		MNLI		SNLI		MNLI	
	In-Dom.	HANS Δ	In-Dom.	HANS Δ	In-Dom.	HANS Δ	In-Dom.	HANS Δ
Baseline (P+H)	87.41	49.84	79.37	51.87	91.50	67.21	87.96	70.25
Hypothesis-Only	70.86	51.10 +1.25	61.15	51.54 -0.33	70.86	50.13 -17.08	61.15	51.73 -18.53
<i>Loss Modification</i>								
Label Smoothing	87.39	49.83 -0.01	79.35	51.88 +0.01	91.64	65.27 -1.94	87.90	69.96 -0.29
Confidence Penalty	87.41	49.84 +0.00	79.37	51.90 +0.03	91.50	66.53 -0.68	87.78	70.65 +0.40
<i>Bias-Model-Based</i>								
Ensemble Debiasing	87.38	50.11 +0.27	79.34	51.51 -0.36	91.25	63.93 -3.28	87.36	67.92 -2.33
Product-of-Experts	87.40	49.66 -0.18	79.36	52.25 +0.38	90.89	65.98 -1.23	87.26	72.12 +1.86
DFL $\gamma=2.0$	87.39	49.98 +0.14	79.35	51.43 -0.44	76.89	61.99 -5.22	81.92	67.63 -2.62
DFL $\gamma=3.0$	87.38	50.11 +0.27	79.33	51.58 -0.29	68.54	51.58 -15.63	75.17	61.72 -8.53
<i>Data Augmentation</i>								
ANLI Augmentation	87.42	49.99 +0.15	79.38	52.11 +0.24	91.52	73.06 +5.85	87.87	73.43 +3.18
Contrastive Aug.	87.40	49.95 +0.10	79.36	52.31 +0.44	91.05	65.19 -2.02	87.58	73.84 +3.59
Aggressive Aug.	87.39	49.90 +0.06	79.34	50.91 -0.96	91.39	67.66 +0.45	87.64	71.62 +1.37

Table 1: In-domain and HANS accuracy (%) with Δ vs. baseline for all 44 settings.

on MNLI, with entailment bias of 99.56% and 94.83% respectively. The premise-irrelevance gap (full model minus Hypothesis-Only) is diagnostic: ELECTRA’s full model performs no better than a model discarding the premise entirely (-1.26 pp SNLI, $+0.33$ pp MNLI), while RoBERTa’s gap is $+17.08$ pp and $+18.52$ pp. ELECTRA learns nothing from the premise; RoBERTa genuinely conditions on it. Hypothesis-Only in-domain scores are identical across architectures as both exploit the same hypothesis-text artifacts.

3.2 ELECTRA-small: Complete Debiasing Failure

On ELECTRA/SNLI, HANS accuracy across all ten debiasing methods spans 49.66%–51.10%, entirely within the baseline CI of [49.28, 50.41]; on ELECTRA/MNLI, all ten fall within [51.30, 52.44]. The coefficient of variation is 0.75% (SNLI) and 0.79% (MNLI): the entire method distribution is tighter than a single 95% CI. Wilcoxon tests confirm no systematic improvement: SNLI $W = 7.0$, $p = 0.070$; MNLI $W = 21.0$, $p = 0.557$. The sole exception (Hypothesis-Only, 51.10%) discards the premise entirely; its ranking first is itself the finding. Aggressive Augmentation on MNLI (50.91%) falls *below* the CI lower bound: the one method that moves ELECTRA does so negatively.

3.3 RoBERTa-base: Real but Inconsistent Effects

RoBERTa shows genuine method sensitivity: CoV is 10.76% on SNLI and 9.42% on MNLI, roughly

$12\text{--}14\times$ higher than ELECTRA’s 0.75% and 0.79%. On SNLI, seven of ten methods hurt HANS accuracy; median delta -1.98 pp, a trend that falls short of conventional significance (Wilcoxon $W = 9.0$, $p = 0.065$). ANLI augmentation is the standout: $+5.85$ pp on SNLI (73.06%) and $+3.18$ pp on MNLI (73.43%), with the largest constituent gains ($+9.26$ pp SNLI, $+5.09$ pp MNLI), preserving in-domain accuracy exactly (91.52% vs. 91.50% baseline). DFL ($\gamma = 3.0$) is the sharpest failure: HANS collapses to 51.58% (-15.63 pp) while in-domain falls from 91.50% to 68.54% (-22.96 pp). Contrastive augmentation gains $+3.59$ pp on MNLI but loses -2.02 pp on SNLI, a 5.6 pp reversal with no change in model, method, or hyperparameters. Spearman $\rho = -0.42$ ($p = 0.20$) for ELECTRA and $\rho = 0.75$ ($p = 0.009$) for RoBERTa: ELECTRA rankings are anti-correlated noise across corpora; RoBERTa rankings are stable and real.

3.4 Robustness-Accuracy Correlation

The $\Delta\text{HANS}\text{--}\Delta\text{in-domain}$ correlation is $r = -0.096$ for ELECTRA, consistent with its flat method distribution, and $r = 0.848$ for RoBERTa. On RoBERTa, the assumed tradeoff inverts: methods that improve HANS preserve in-domain accuracy; methods that aggressively suppress bias (DFL, Ensemble Debiasing) lose on both axes. ANLI augmentation ($+5.85$ pp HANS, zero in-domain cost) and DFL $\gamma = 3.0$ (-15.63 pp HANS, -22.96 pp in-domain) define the poles.

3.5 Per-Heuristic Breakdown (RoBERTa-base)

Table 2 shows HANS per-heuristic accuracy for RoBERTa-base across both corpora. ELECTRA per-heuristic results are omitted: all values cluster within 49–56% across all methods and heuristics, consistent with the flat CoV reported in §3. The breakdown reveals that aggregate HANS accuracy obscures qualitatively distinct failure modes.

Method	SNLI			MNLI		
	Lex	Sub	Con	Lex	Sub	Con
Baseline	82.92	61.18	57.53	81.95	64.00	64.81
Label Smooth.	82.13	59.10	54.58	81.53	63.91	64.45
Conf. Penalty	82.63	60.39	56.58	81.87	64.59	65.49
Ensemble	76.74	58.30	56.75	78.64	61.75	63.38
PoE	85.55	57.52	54.87	85.18	65.17	66.00
DFL $\gamma=2.0$	75.86	56.31	53.79	75.46	61.69	65.74
DFL $\gamma=3.0$	51.98	50.94	51.83	63.29	60.43	61.45
ANLI Aug.	87.88	64.52	66.79	84.14	66.26	69.90
Contrastive	79.41	57.38	58.79	86.84	65.34	69.34
Aggressive	85.27	60.26	57.45	81.37	64.18	69.32
Hyp-Only	49.88	50.27	50.23	50.08	55.09	50.01

Table 2: HANS per-heuristic accuracy (%) for RoBERTa-base. Lex = lexical overlap, Sub = subsequence, Con = constituent.

4 Analysis and Discussion

The capacity hypothesis. Bias-model-based methods require a functional feedback loop: the bias model flags artifact-reliant examples so the main model can down-weight them. On ELECTRA, this loop fails at the source. The hypothesis-only bias model achieves $\sim 51\%$ on HANS, near chance, so the debiasing loss reduces to plain cross-entropy; the machinery runs but steers with noise. Bias-free methods fail too: label smoothing and confidence penalty cannot inject cross-sentence reasoning capacity the architecture lacks; augmentation cannot force premise-sensitivity into a model that consistently ignores the premise. The premise-irrelevance gap (-1.26 pp on SNLI) makes this structural: the capacity ceiling is set before any intervention begins.

Pretraining as a confound. The gap is not solely parameter count. ELECTRA’s Replaced-Token-Detection optimises for local token-level prediction; RoBERTa’s Masked Language Modelling requires broader cross-sentence representations. This inductive bias difference likely restricts ELECTRA’s ability to build premise-hypothesis repre-

sentations independent of size. A method validated exclusively on MLM-pretrained encoders cannot be assumed to transfer to RTD-pretrained models, an unstated precondition of almost every published debiasing result.

The publication selection effect. The approximately $14\times$ CoV difference (0.75% vs. 10.76%) exposes the problem precisely. A researcher running only RoBERTa observes real sensitivity and publishes a working method. A researcher running only ELECTRA observes flat noise and has nothing to publish. The published record consists of the first kind of result. Same codebase, same seed, same data: one architecture moves, the other does not. What the field calls method effectiveness is the interaction of method and architecture, inseparable without cross-architecture controls.

Dataset-method interactions. The 5.6 pp contrastive augmentation reversal across SNLI and MNLI reflects a systematic interaction between generated examples and corpus artifact distributions. SNLI’s densely patterned crowdsourced hypotheses may cause rule-based contradictions to reinforce existing artifacts rather than disrupt them; MNLI’s multi-genre hypotheses present a more varied artifact landscape the same contradictions successfully target. ANLI augmentation’s constituent gains ($+9.26$ pp SNLI, $+5.09$ pp MNLI) follow the same logic: adversarially collected examples target the failure modes most prevalent in the corpus they were built against. These interactions are invisible to single-corpus evaluation.

The positive robustness-accuracy correlation. ANLI augmentation introduces constructive gradient signal where heuristics misfire, expanding the decision boundary without disrupting in-distribution learning. DFL aggressively down-weights artifact-aligned examples; when the model cannot find alternative generalisable features, it destroys both axes simultaneously. Within the scope of the artifacts HANS captures, the robustness-accuracy tradeoff is not axiomatic; it is a property of overly aggressive methods, not of debiasing in general. We do not claim this extends to bias patterns HANS does not test.

5 Conclusion

We set out to test an assumption the NLI debiasing literature has never examined: that a method

validated on one architecture and one corpus generalises beyond it. Across 44 controlled experiments it does not. Debiasing method effectiveness is not a property of the method in isolation; it is jointly determined by the method and the architecture receiving it, and the two cannot be separated without exactly the cross-architecture design the field has been skipping.

The practical consequence is a publication selection effect: a researcher who happens to run RoBERTa observes real, if inconsistent, method sensitivity and has a paper; a researcher who happens to run ELECTRA observes flat noise and has nothing to publish, even on an identical codebase with an identical seed. What the field records as method effectiveness is this selection artefact as much as it is genuine progress. The clearest evidence against a simple robustness-accuracy trade-off is ANLI augmentation, which improves HANS accuracy at essentially zero in-domain cost, set against DFL at its aggressive setting, which damages both simultaneously; the tradeoff, where it appears at all, is a symptom of method aggressiveness rather than an inherent property of debiasing.

We do not read these results as a reason to abandon debiasing research, but as a case for conducting it with the rigour its evaluation infrastructure has so far lacked: cross-architecture validation on at least two structurally distinct models, Wilson confidence intervals with paired significance tests, model capacity reported as a moderating variable, and HANS per-heuristic breakdowns rather than a single aggregate number. Without these controls, a positive result is not wrong so much as impossible to interpret.

Limitations.

ELECTRA-small and RoBERTa-base differ in both parameter count (14M vs. 125M) and pretraining objective (Replaced-Token-Detection vs. Masked Language Modelling), so the two factors are confounded in our design: we cannot conclusively attribute ELECTRA’s failure to representational capacity alone rather than to properties of RTD pretraining specifically. We favour the capacity explanation because it is consistent with the premise-irrelevance gap, a within-model diagnostic that does not depend on comparing across architectures, but isolating size from pretraining objective would require a small MLM-pretrained model (e.g., BERT-tiny) run under the same protocol, which we

leave to future work. Results may not generalise to larger architectures (RoBERTa-large, DeBERTa) or non-English NLI. All runs use a single random seed; multiple seeds would further strengthen stability claims. The contrastive pipeline is rule-based, and HANS covers three heuristics but does not exhaust NLI artifact space.

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](#). *Preprint*, arXiv:2003.10555.
- 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 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.
- Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. [End-to-end bias mitigation by modelling biases in corpora](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.
- 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*.
- 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.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *ICLR Workshop*. ArXiv:1701.06548.
- 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, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.