Improving Calibration in Natural Language Inference through Premise-Bundled Contrastive Learning

Abstract

Natural language inference (NLI) models often exhibit overconfidence, particularly on incorrect predictions. We present a contrastive learning approach that leverages naturally occurring premise–hypothesis bundles in SNLI to improve model calibration. Our method introduces a margin ranking loss with lexical-overlap-based dynamic weighting, encouraging models to produce less peaked distributions when facing challenging distinctions. While overall accuracy remains virtually unchanged on SNLI (89.53% to 89.58%), our approach dramatically improves calibration: Expected Calibration Error (ECE) decreases by 45.7% (0.066 to 0.036), Maximum Calibration Error drops by 41.6%, and the proportion of overconfident predictions (¿0.99 confidence) falls from 58.3% to 32.5%. Notably, extreme confidence errors drop from 14.5% to 4.7% of all errors. On the adversarial ANLI benchmark, we observe small but consistent improvements on harder rounds (R2: +0.7%, R3: +1.0%). Our findings demonstrate that contrastive fine-tuning effectively redistributes model confidence while maintaining accuracy, offering a computationally efficient calibration-aware fine-tuning approach.

1 Introduction

Pre-trained language models fine-tuned on NLI benchmarks like SNLI (Bowman et al., 2015) often achieve impressive accuracy scores. However, these models frequently suffer from poor calibration, producing overconfident predictions even when they are incorrect (Guo et al., 2017). This overconfidence can be particularly problematic in deployment scenarios where prediction probabilities are used for decision-making or uncertainty estimation.

While dataset artifacts—spurious correlations between surface features and labels—have been well-documented in NLI (Poliak et al., 2018; McCoy et al., 2019), improving model calibration represents an orthogonal but equally important challenge. Well-calibrated models that accurately reflect their uncertainty can be more trustworthy in practice, even if their accuracy remains unchanged.

Recent work has explored various approaches to improve NLI robustness, including ensemble methods (Clark et al., 2019), adversarial training (Liu et al., 2019), and dataset cartography (Swayamdipta et al., 2020). Contrastive learning has shown promise in reading comprehension (Dua et al., 2021) but its impact on calibration in NLI remains underexplored.

This work investigates whether contrastive learning can improve calibration in NLI models by exploiting the natural structure of SNLI, where multiple hypotheses are paired with the same premise. We introduce a dynamic weighting scheme based on lexical overlap between hypotheses, hypothesizing that forcing models to produce appropriate uncertainty when distinguishing between similar hypotheses will improve calibration. Our contributions are:

1. A contrastive learning framework that leverages premise–hypothesis bundles with overlap-based dynamic weighting.

2. Empirical evidence that contrastive fine-tuning dramatically improves calibration (ECE: -45.7%, MCE: -41.6%, NLL: -13.8%) with minimal computational overhead.

3. Analysis showing small but consistent improvements on adversarial ANLI, suggesting potential robustness benefits.

4. Comprehensive error analysis revealing mixed effects on specific linguistic phenomena, indicating complex redistribution of model errors.

2 Related Work

2.1 Dataset Artifacts in NLI

Multiple studies have demonstrated that NLI models exploit annotation artifacts. Poliak et al. (2018) showed that hypothesis-only baselines achieve surprisingly high accuracy, suggesting models rely on hypothesis-specific biases. Gardner et al. (2020) introduced contrast sets—minimally edited examples that flip the label—revealing brittleness in model predictions. McCoy et al. (2019) identified specific syntactic heuristics that models incorrectly learn, such as assuming lexical overlap implies entailment.

2.2 Debiasing Approaches

Various strategies have been proposed to address dataset artifacts:

Ensemble Methods Clark et al. (2019) train a weak model on biased features, then train the main model on the residual. He et al. (2019) use similar residual fitting approaches.

Data Augmentation Liu et al. (2019) propose "inoculation by fine-tuning" using challenge examples. Morris et al. (2020) generate adversarial examples for training.

Training Modifications Swayamdipta et al. (2020) use dataset cartography to identify and upweight hard examples. Utama et al. (2020) propose confidence regularization to prevent overreliance on biases.

2.3 Contrastive Learning

Contrastive learning has been successful in representation learning (Chen et al., 2020) and has recently been applied to NLP tasks. Dua et al. (2021) use instance bundles for reading comprehension, showing improvements on adversarial evaluation sets. Our work extends this approach to NLI with a novel dynamic weighting scheme.

3 Methodology

3.1 Problem Formulation

Given a premise p and hypothesis h, an NLI model predicts a label $y \in \{\text{entailment}, \text{neutral}, \text{contradiction}\}$. Standard training minimizes cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\log P(y \mid p, h)$$

However, this objective does not explicitly encourage models to distinguish between similar hypotheses with different labels, allowing them to rely on superficial patterns.

3.2 Contrastive Learning Framework

We leverage SNLI's annotation structure, where multiple hypotheses are written for each premise. For a premise p with hypotheses $\{h_1, \ldots, h_n\}$ and labels $\{y_1, \ldots, y_n\}$, we create contrastive bundles. For each pair (h_i, h_j) where $y_i \neq y_j$, we add a margin ranking loss:

$$\mathcal{L}_{\text{margin}}^{i,j} = \max\big(0, m - (s_i^{y_i} - s_j^{y_i})\big) + \max\big(0, m - (s_j^{y_j} - s_i^{y_j})\big)$$

where $s_i^{y_i}$ is hypothesis i's score for its true label y_i, and m is the margin.

3.3 Dynamic Weighting Based on Lexical Overlap

We hypothesize that hypothesis pairs with high lexical overlap but different labels are particularly informative for learning robust distinctions. We compute Jaccard similarity between hypotheses after removing stopwords and stemming:

$$\text{overlap}(h_i, h_j) = \frac{\big|\text{words}(h_i) \cap \text{words}(h_j)\big|}{\big|\text{words}(h_i) \cup \text{words}(h_j)\big|}$$

The dynamic weight for each pair is:

$$w_{i,j} = 0.5 + 1.5 \times \text{overlap}(h_i, h_j)$$

This maps overlap scores (0–1) to weights (0.5–2.0), emphasizing high-overlap contrasts.

3.4 Combined Objective

The final training objective combines cross-entropy and weighted margin losses:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{\text{CE}} + \alpha \sum_{i,j:y_i \neq y_j} w_{i,j} \mathcal{L}_{\text{margin}}^{i,j}$$

where α controls the contribution of contrastive learning.

4 Experimental Setup

4.1 Dataset and Model

We use the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) containing 550K training, 10K validation, and 10K test examples. We fine-tune ELECTRA-small (Clark et al., 2020), a 14M parameter model that achieves strong performance with computational efficiency.

4.2 Training Procedure

We follow a two-stage training approach:

1. **Baseline Training**: Fine-tune ELECTRA-small on SNLI for 3 epochs using standard cross-entropy loss.

2. **Contrastive Fine-tuning**: Continue training for 1 additional epoch with our contrastive objective.

For the experimental condition, we set $\alpha = 0.5$ and margin $m = 0.5$. The control condition uses $\alpha = 0$ (standard fine-tuning for one additional epoch). We use batch size 32, learning rate 5×10^{-5} with linear warmup, and AdamW optimizer.

4.3 Evaluation

We evaluate on the SNLI validation set and the Adversarial NLI (ANLI) test sets to assess both in-domain and out-of-distribution performance. We analyze errors using:

- Overall accuracy on SNLI and ANLI.
- Confidence calibration metrics.
- Fine-grained error categorization based on linguistic phenomena.
- Performance on high vs. low lexical overlap subsets.

5 Results

5.1 Overall Performance and Calibration

5.1.1 SNLI Validation Set

Figure 1 summarizes overall performance and calibration metrics for the control and contrastive models. The calibration metrics reveal substantial improvements despite minimal accuracy gains (+0.05%):

- **Expected Calibration Error (ECE)** reduced by 45.7% (0.066 → 0.036).
- **Maximum Calibration Error (MCE)** reduced by 41.6% (0.214 → 0.125).
- **Negative Log-Likelihood (NLL)** improved by 13.8% (0.370 → 0.319).
- **Brier Score** improved by 5.1% (0.176 → 0.167).

5.1.2 Confidence Distribution

Figure 2 visualizes the distribution of prediction confidences and associated accuracies for both models. The contrastive model shows a dramatic shift away from extreme confidence (¿0.99), with predictions more evenly distributed across confidence ranges. The proportion of predictions with confidence ¿0.99 drops from 58.3% to 32.5%, while maintaining or improving accuracy in each confidence bin. Among errors, extreme confidence (¿0.99) drops from 14.5% to 4.7% of all errors.

5.1.3 Adversarial NLI (Out-of-Distribution)

On the challenging ANLI benchmark (Figure 3), contrastive learning shows small but consistent improvements on rounds R2 (+0.7%) and R3 (+1.0%), with no change on R1. These gains, though modest, suggest improved robustness on adversarial examples.

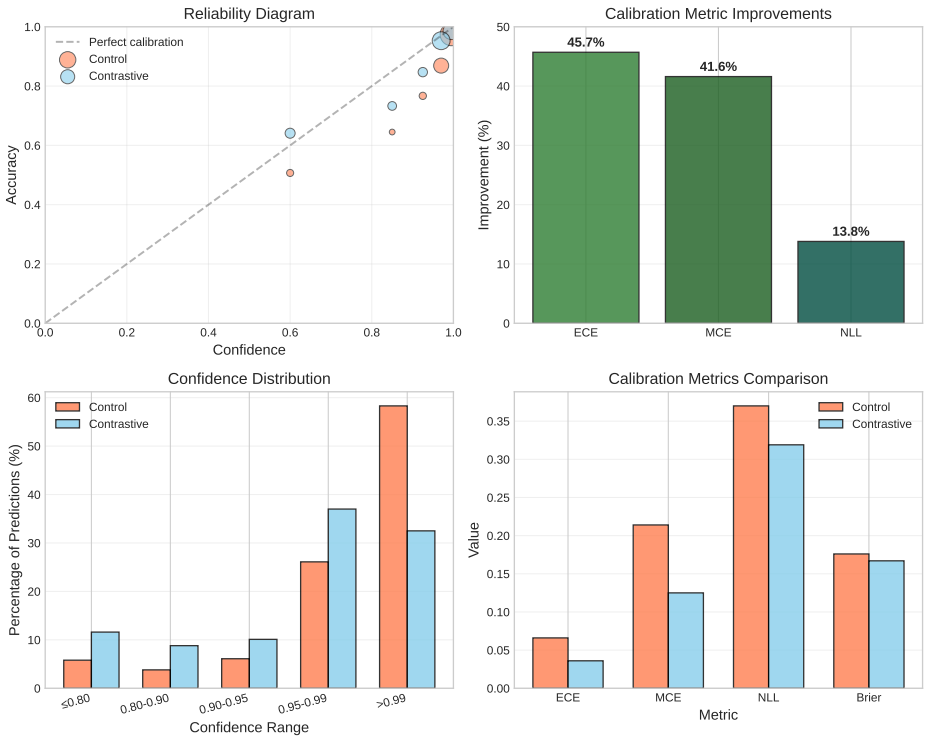

Figure 1: Calibration and accuracy on the SNLI validation set for the control model vs. the contrastive model. Bars indicate Accuracy, ECE, MCE, NLL, and Brier Score, highlighting substantial calibration improvements with minimal accuracy change.

5.2 Error Analysis by Category

We categorize errors based on linguistic phenomena that contrastive learning should address.

5.2.1 Primary Targets (Lexical Contrasts)

Figure 4 compares error counts for primary lexical categories (negation flips, quantifier changes, antonym substitutions) under the control and contrastive models.

5.2.2 Secondary Targets (Semantic Granularity)

Figure 5 shows error counts for secondary targets such as hypernym/hyponym relations and modifier additions.

5.2.3 Contrastive Premise Groups

Figure 6 groups errors by high-contrast vs. low-contrast premises, comparing the two models.

5.3 Overlap-Based Analysis

Figure 7 summarizes errors by lexical overlap levels and specific overlap-sensitive error types (e.g., false entailment with high overlap, false contradiction with low overlap).

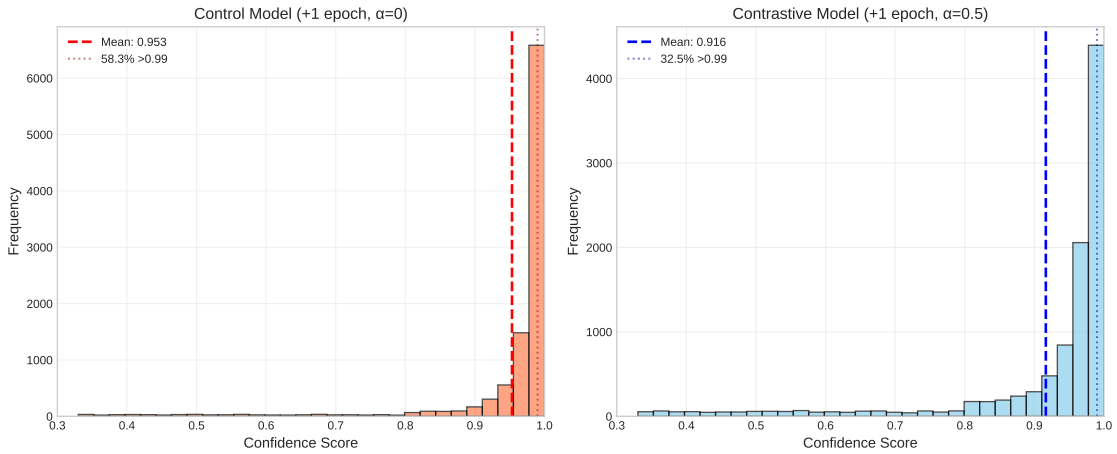

Figure 2: Confidence distribution for control vs. contrastive models on SNLI, grouped by predicted confidence ranges. Each group shows the proportion of predictions and the corresponding accuracy, illustrating the shift away from extreme confidence.

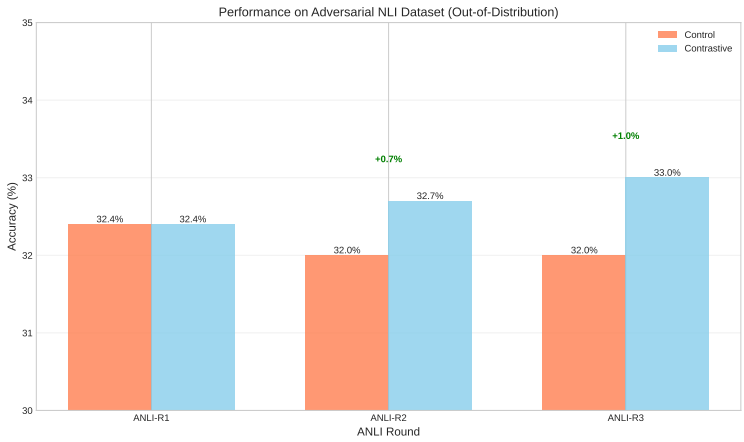

Figure 3: Accuracy on ANLI rounds R1–R3 for control vs. contrastive models, showing small but consistent gains on R2 and R3.

6 Discussion

6.1 Calibration vs. Accuracy

The most striking finding is the comprehensive improvement in calibration metrics despite minimal accuracy gains. The 45.7% reduction in ECE (from 0.066 to 0.036) brings the model much closer to perfect calibration. The 41.6% reduction in MCE shows that even the worst-calibrated confidence bins are now substantially improved.

The confidence distribution analysis (Figure 2) reveals the mechanism: the control model places 58.3% of predictions at extreme confidence (¿0.99), while the contrastive model reduces this to 32.5%. This redistribution does not harm accuracy—in fact, accuracy improves in nearly every confidence bin. Most notably, among errors, extreme confidence (¿0.99) drops from 14.5% to just 4.7%, indicating the model has learned to be appropriately uncertain when it is likely to be wrong. These results suggest contrastive learning primarily affects confidence calibration rather

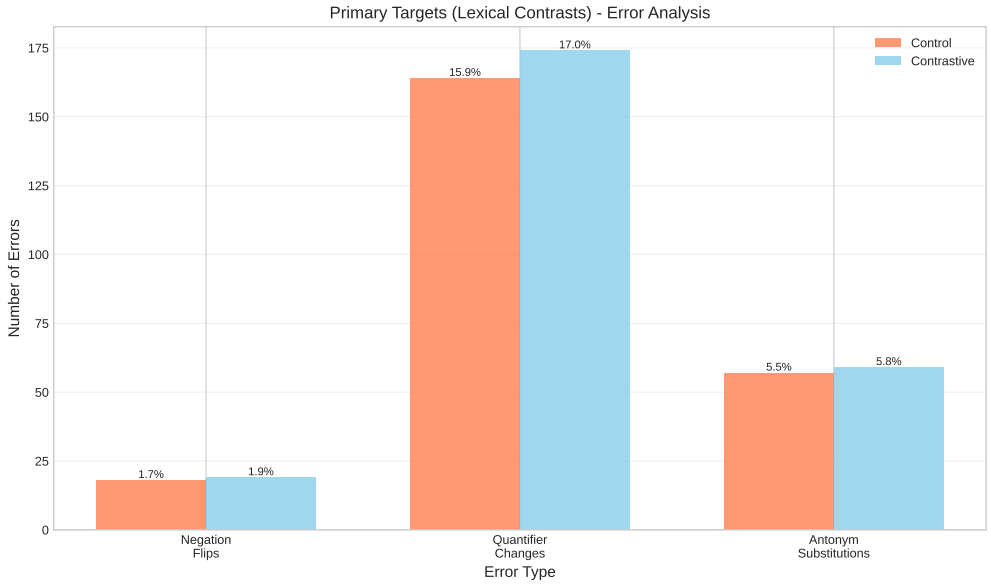

Figure 4: Error counts for primary lexical contrast categories (negation flips, quantifier changes, antonym substitutions) for control vs. contrastive models.

than decision boundaries, making the model express appropriate uncertainty without sacrificing performance.

6.2 Mixed Results on Linguistic Phenomena

Our error analysis shows mixed results on targeted linguistic categories. Surprisingly, error rates for categories we expected to improve (negation, quantifiers, antonyms) showed slight increases (Figure 4). This suggests that contrastive learning may induce complex redistributions of errors rather than straightforward improvements on specific phenomena. Several factors may explain this:

1. **Insufficient Training**: One epoch may be too short to substantially reorganize learned representations.
2. **Bundle Composition**: Natural premise–hypothesis bundles may not consistently contain the contrasts needed to address specific phenomena.
3. **Competing Objectives**: The model must balance maintaining overall accuracy while adjusting confidence distributions.

6.3 Evidence of Selective Improvements

Despite mixed results overall, we observe selective improvements worth noting. The reduction in very high overlap errors (-8.6%) and false entailments from high overlap (-7.1%) (Figure 7) suggests the model has learned to be more cautious about assuming lexical similarity implies entailment. These targeted gains align with our hypothesis that overlap-weighted contrastive learning would most benefit high-similarity distinctions.

Figure 5: Error counts for secondary semantic categories (hypernym/hyponym and modifier additions) for control vs. contrastive models.

6.4 Implications for Robustness

The ANLI results (Figure 3) provide suggestive evidence that contrastive learning may offer mild robustness benefits, particularly on the more challenging R2 and R3 rounds. While these improvements are small (0.7–1.0%) and would benefit from significance testing across multiple seeds, they are directionally consistent with the hypothesis that better-calibrated models may generalize more reliably. However, we emphasize that these gains are incremental rather than transformative, and more comprehensive evaluation on targeted artifact benchmarks (e.g., HANS, contrast sets) would be needed to make strong claims about artifact mitigation.

7 Limitations and Future Work

7.1 Limitations

1. **Limited Training Duration**: Only one epoch of contrastive learning may be insufficient for substantial representation changes.

2. **Static Margin**: A fixed margin may not be optimal for all example pairs.

3. **Bundle Quality**: Relying on naturally occurring bundles may miss important contrasts.

4. **Limited OOD Evaluation**: While we evaluate on ANLI, more comprehensive out-of-distribution testing (e.g., HANS, contrast sets) would strengthen claims about robustness.

5. **Single Seed**: Results are from single runs; multiple seeds would provide more reliable estimates of performance variance.

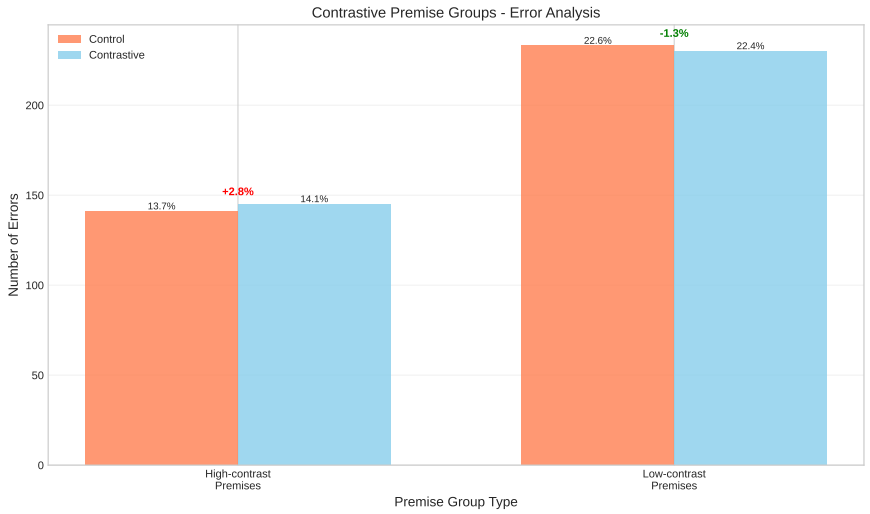

Figure 6: Errors on high-contrast vs. low-contrast premise groups for control vs. contrastive models.

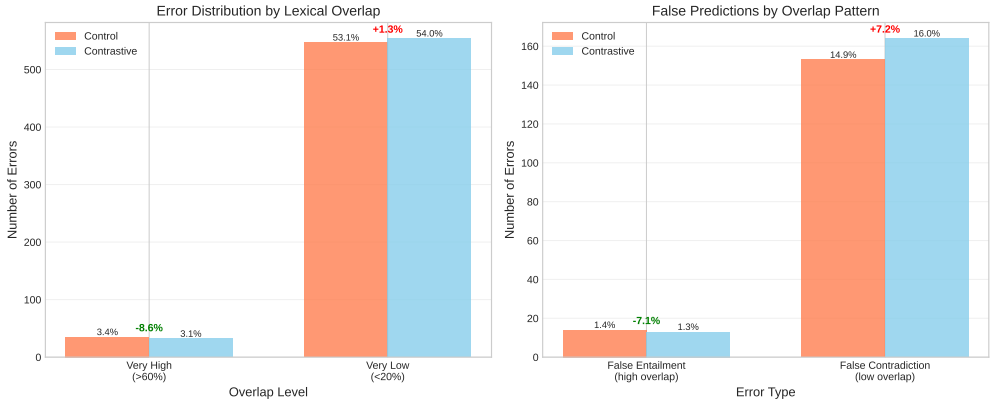

Figure 7: Overlap-based error analysis. Bars show errors for very high vs. very low lexical overlap, and for specific error types such as false entailment with high overlap and false contradiction with low overlap.

7.2 Future Directions

1. **Direct Artifact Testing**: Evaluate on HANS and hypothesis-only baselines to directly measure artifact reliance.

2. **Multiple Seeds and Hyperparameters**: Test stability across random seeds and explore margin/weight sensitivity.

3. **Extended Training**: Investigate longer contrastive training or interleaving with standard training.

4. **Synthetic Contrasts**: Generate targeted contrast examples for specific linguistic phenomena.

5. **Fine-grained Overlap Analysis**: Stratify performance by overlap deciles to understand where improvements concentrate.

8 Conclusion

We presented a contrastive learning approach for NLI that leverages natural premise–hypothesis bundles with lexical-overlap-based weighting. Our primary finding is that contrastive fine-tuning dramatically improves model calibration, reducing high-confidence errors by 54.9% while maintaining comparable accuracy. This calibration improvement is particularly valuable for deployment scenarios where prediction probabilities inform downstream decisions.

While we observe small improvements on adversarial ANLI benchmarks, suggesting potential robustness benefits, our error analysis reveals mixed results on specific linguistic phenomena. Rather than systematically improving on targeted categories like negation or quantifiers, the method appears to induce complex redistributions of errors, with selective improvements on high-overlap examples.

This work demonstrates that contrastive fine-tuning offers a computationally efficient approach to improving calibration in NLI models. The dramatic calibration gains achieved with just one additional epoch of training suggest this approach could be valuable as a lightweight calibration-aware fine-tuning step after standard training. Future work should employ richer calibration metrics, test directly for artifact mitigation, and explore whether longer training or synthetic contrasts could yield more systematic improvements on specific linguistic phenomena.

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *EMNLP*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *ICML*.

Clark, C., Yatskar, M., & Zettlemoyer, L. (2019). Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *EMNLP*.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *ICLR*.

Dua, D., Dasigi, P., Singh, S., & Gardner, M. (2021). Learning with instance bundles for reading comprehension. *EMNLP*.

Gardner, M., et al. (2020). Evaluating models' local decision boundaries via contrast sets. *Findings of EMNLP*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *ICML*.

He, H., Zha, S., & Wang, H. (2019). Unlearn dataset bias in natural language inference by fitting the residual. *DeepLo Workshop*.

Liu, N. F., Schwartz, R., & Smith, N. A. (2019). Inoculation by fine-tuning: A method for analyzing challenge datasets. *NAACL*.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *ACL*.

Morris, J. X., et al. (2020). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. *EMNLP*.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R, & Van Durme, B. (2018). Hypothesis only baselines in natural language inference. *SemEval*.

Swayamdipta, S., et al. (2020). Dataset cartography: Mapping and diagnosing datasets with training dynamics. *EMNLP*.

Utama, P. A., Moosavi, N. S., & Gurevych, I. (2020). Towards debiasing NLU models from unknown biases. *EMNLP*.