Improving NLI Reliability through Contrastive Calibration and Adversarial Fine-tuning

December 2025

Abstract

Despite strong performance on standard benchmarks, transformer-based NLI systems suffer from systematic overconfidence and vulnerability to distribution shifts. We investigate two lightweight fine-tuning methods using ELECTRA-small: first, a contrastive approach leveraging naturally grouped premise-hypothesis pairs in SNLI with dynamic weighting based on lexical similarity; second, an adversarial strategy incorporating ANLI examples during training. Our contrastive method maintains accuracy (89.53% → 89.58%) while achieving substantial calibration improvements—ECE decreases 45.7% (0.066 → 0.036), MCE drops 41.6%, and predictions exceeding 99% confidence fall from 58.3% to 32.5%. Critically, highly confident errors decrease from 14.5% to 4.7% of mistakes. The adversarial approach, training jointly on SNLI and ANLI for one epoch, increases combined ANLI performance by 10.8 points (34% relative gain) without degrading SNLI accuracy. These complementary strategies target different failure modes: contrastive learning redistributes model confidence to better reflect uncertainty, while adversarial training directly addresses specific reasoning failures. Together, they offer computationally efficient paths toward more reliable NLI systems.

1. Introduction

Natural language inference (NLI) evaluates whether machines can determine logical relationships—entailment, contradiction, or neutrality—between sentence pairs (Bowman et al., 2015). Although transformer models achieve impressive benchmark results on SNLI, research has uncovered concerning reliance on superficial patterns. Several studies demonstrate these limitations: models trained without premises still perform surprisingly well (Poliak et al., 2018), lexical similarity between sentences heavily influences predictions (McCoy et al., 2019), and small meaning-preserving changes trigger incorrect outputs (Gardner et al., 2020). Such vulnerabilities question whether models truly understand language or simply exploit statistical regularities.

Two reliability issues are particularly problematic in deployment:

1. **Overconfidence and poor calibration.** NLI models frequently assign near-certainty to incorrect predictions (Guo et al., 2017). In applications where predicted probabilities inform downstream decisions or risk assessments, miscalibration can be as harmful as low accuracy.
2. **Brittleness to adversarial examples.** Adversarial challenge sets like ANLI (Nie et al., 2020) employ iterative annotation processes where humans craft examples that fool current models, targeting specific weaknesses (lexical substitutions, negation, world knowledge). Models with high SNLI accuracy frequently drop to chance-level performance on ANLI.

Recent work has proposed several strategies to address these failures. To combat artifacts and shortcut learning, researchers have explored debiasing via ensembles (Clark et al., 2019; He et al., 2019), data augmentation and adversarial training (Liu et al., 2019; Morris et al., 2020), and instance reweighting based on dataset cartography (Swayamdipta et al., 2020) or confidence regularization (Utama et al., 2020). In parallel, calibration-specific approaches aim to align predicted probabilities with empirical correctness, using techniques such as temperature scaling (Guo et al., 2017) or calibration-aware training objectives. Contrastive learning, which encourages models to distinguish between similar but label-differing instances, has shown

promise in reading comprehension and representation learning (Chen et al., 2020; Dua et al., 2021), but its calibration effects in NLI remain underexplored.

In this project we investigate two complementary fine-tuning strategies on a shared ELECTRA-small backbone (Clark et al., 2020) that directly target these reliability gaps:

- **A premise-bundled contrastive learning objective** on SNLI, which adds an overlap-weighted margin ranking loss over hypotheses sharing the same premise but having different labels. This objective is designed to improve calibration by forcing the model to express appropriate uncertainty on subtle lexical and semantic contrasts.
- **An adversarial fine-tuning regime** that mixes ANLI with SNLI for a brief additional epoch. This setup uses adversarially constructed ANLI examples to "correct" the SNLI-trained model, with the goal of improving performance on ANLI while preserving SNLI accuracy.

We frame these as two halves of a broader question: can we improve both calibration and adversarial robustness of an NLI model using lightweight, post-hoc fine-tuning strategies that minimally disturb its strong in-domain performance?

Our main contributions are:

1. **Contrastive calibration for NLI.** We propose a contrastive fine-tuning framework that leverages naturally occurring premise-hypothesis bundles in SNLI, with dynamic weights based on lexical overlap. We show that a single contrastive epoch yields large gains in calibration metrics with essentially unchanged SNLI accuracy.
2. **Adversarial robustness via ANLI fine-tuning.** We establish a strong SNLI ELECTRA-small baseline and compare an additional SNLI epoch (control) to a joint SNLI+ANLI fine-tuning regime. The adversarial run improves ANLI accuracy by +10.8 absolute points while preserving SNLI performance.
3. **Joint perspective on reliability.** By analyzing calibration and adversarial robustness together, we highlight how contrastive and adversarial fine-tuning tackle different failure modes: contrastive learning primarily reshapes confidence distributions, whereas ANLI fine-tuning primarily improves correctness on hard adversarial examples.
4. **Qualitative and categorical error analysis.** We present fine-grained error breakdowns across linguistic phenomena (negation, quantifiers, lexical overlap, world knowledge) and discuss where each strategy helps or fails, revealing complex redistributions of errors rather than uniform improvements.

---

2. Background and Related Work

2.1 Dataset Artifacts and Shortcut Learning in NLI

Studies have uncovered critical flaws in NLI model behavior. Standard training often leads models to exploit statistical shortcuts instead of learning robust inference. Poliak et al. (2018) showed that models trained exclusively on hypotheses—ignoring premises entirely—still achieve substantial accuracy, indicating reliance on annotation biases. McCoy et al. (2019) documented specific heuristics models adopt, including treating lexical similarity as evidence for entailment. Gardner et al. (2020) demonstrated fragility through small meaning-preserving modifications that nonetheless trigger errors, confirming dependence on shallow patterns.

To mitigate artifacts, several lines of work modify training data or objectives:

- **Ensemble and residual debiasing.** Clark et al. (2019) and He et al. (2019) train a "biased" model on spurious features and encourage a main model to focus on residual information.
- **Data augmentation and adversarial training.** Liu et al. (2019) propose "inoculation by fine-tuning", adding targeted challenge examples to training. Morris et al. (2020) generate adversarial examples via meaning-preserving transformations to expose model weaknesses.
- **Instance reweighting and calibration-aware objectives.** Swayamdipta et al. (2020) introduce dataset cartography, identifying "hard" and "ambiguous" instances to guide reweighting. Utama et al. (2020) propose confidence regularization to prevent overreliance on biased cues.

Our work connects to this line by using premise-bundled contrastive pairs and adversarially constructed ANLI examples as targeted training signals against shortcut behavior.

2.2 Calibration of Neural Classifiers

Calibration refers to the alignment between a model's predicted confidence and its actual accuracy—a well-calibrated model should be correct approximately 80% of the time when it predicts with 80% confidence (Guo et al., 2017). Contemporary neural networks, despite their strong performance, frequently suffer from systematic overconfidence. This miscalibration becomes especially pronounced with standard training practices involving cross-entropy optimization and various regularization techniques. Such confidence misalignment poses serious risks in practical applications where probability estimates guide critical decisions or determine system behavior thresholds.

Standard post-hoc calibration methods (e.g., temperature scaling) adjust predicted logits without changing the model's decision boundaries. Training-time approaches instead modify objectives or architectures to encourage calibrated behavior, for example by adding regularizers that penalize miscalibrated confidence distributions. In NLI, however, most prior work has focused on accuracy and artifact mitigation rather than calibration metrics such as Expected Calibration Error (ECE), Maximum Calibration Error (MCE), Negative Log-Likelihood (NLL), or Brier score.

Our contrastive fine-tuning approach falls into this training-time calibration category. By explicitly contrasting hypotheses that share a premise but have different labels—and by weighting contrasts based on lexical overlap—we encourage the model to avoid unwarranted extreme confidence when distinctions are subtle.

2.3 ELECTRA and Adversarial Challenge Sets

ELECTRA (Clark et al., 2020) employs a discriminative pre-training strategy distinct from masked language modeling. The model learns to detect tokens that a generator has substituted within input sequences, training as a binary classifier for each position. This approach yields efficient representations particularly suited for classification. We selected ELECTRA-small (14M parameters) for its computational accessibility while maintaining competitive accuracy.

Multiple adversarial benchmarks evaluate model robustness beyond standard test sets. ANLI (Nie et al., 2020) stands out through its iterative collection process: annotators create examples that challenge model predictions across successive rounds. The dataset comprises three stages (A1-A3) spanning diverse sources—Wikipedia, news articles, fiction, dialogue, and instructions—with progressively harder examples. SNLI-trained models typically perform poorly on ANLI (~33%) accuracy), highlighting severe generalization failures.

Prior work has shown that fine-tuning on adversarial data can improve robustness, but such gains may come at the expense of in-domain performance or may require complex training curricula. Our adversarial fine-tuning experiments adopt a simple regime—one additional epoch on concatenated SNLI+ANLI—to test how much robustness can be gained without harming SNLI accuracy.

2.4 Contrastive Learning in NLP

Contrastive learning has proven effective for representation learning (Chen et al., 2020) with growing adoption in NLP. For reading comprehension, Dua et al. (2021) group questions that share passages but require different answers, then apply contrastive losses to strengthen discrimination between similar examples. This bundling strategy demonstrates gains on challenging evaluation benchmarks.

Our contrastive NLI approach follows a similar bundle-based philosophy: we treat all hypotheses associated with a given SNLI premise as a bundle and impose margin-based ranking constraints on label scores for hypothesis pairs with different gold labels. Our main novelty lies in the lexical-overlap-based dynamic weighting, which prioritizes high-overlap contrasts thought to be especially challenging and calibration-relevant.

Training Configuration and Evaluation

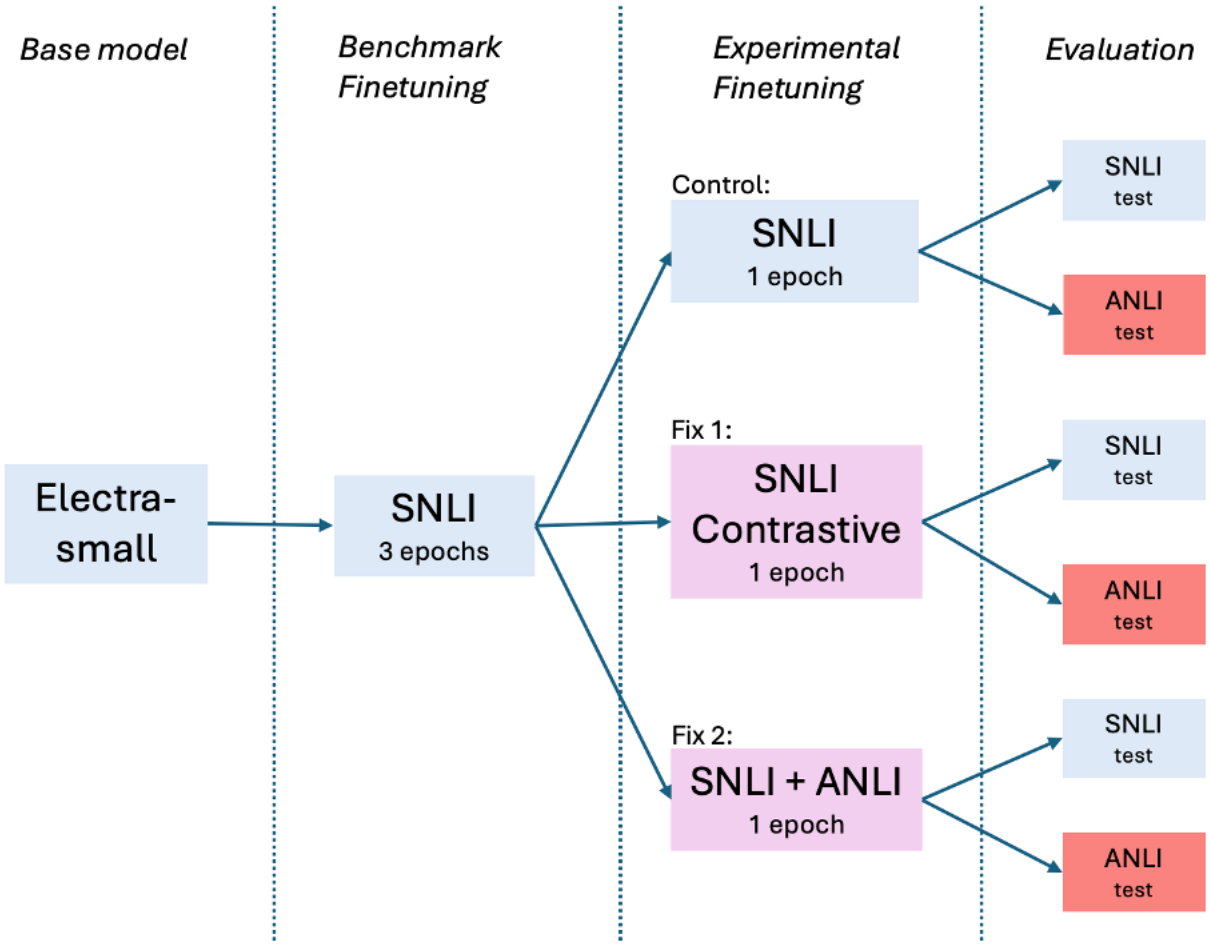

Figure 1: Training flowchart for both contrastive and adversarial fine-tuning approaches

Hyperparameters	**Evaluation Metrics**
Batch size: 8	Accuracy (SNLI, ANLI per round)
Learning rate: 5e-5 with linear warmup	Calibration (ECE, MCE, NLL, Brier)
Optimizer: AdamW	Error categories by linguistic phenomena
Max sequence length: 128 tokens	Performance by lexical overlap level
Weight decay: 0.0	
Max gradient norm: 1.0	

Part I: Contrastive Learning for Calibration

3. Contrastive Methodology

3.1 Problem Formulation

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

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

However, this objective doesn't 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 $\{h_1, ..., h_n\}$ and labels $\{y_1, ..., y_n\}$, we create contrastive bundles. For each pair (h_i, h_j) where $y_i \neq y_j$, we add a margin ranking loss:

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

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:

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

The dynamic weight for each pair is:

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

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:

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

where α controls the contribution of contrastive learning.

4. Contrastive 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

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).

5. Contrastive Results

5.1 Overall Performance and Calibration

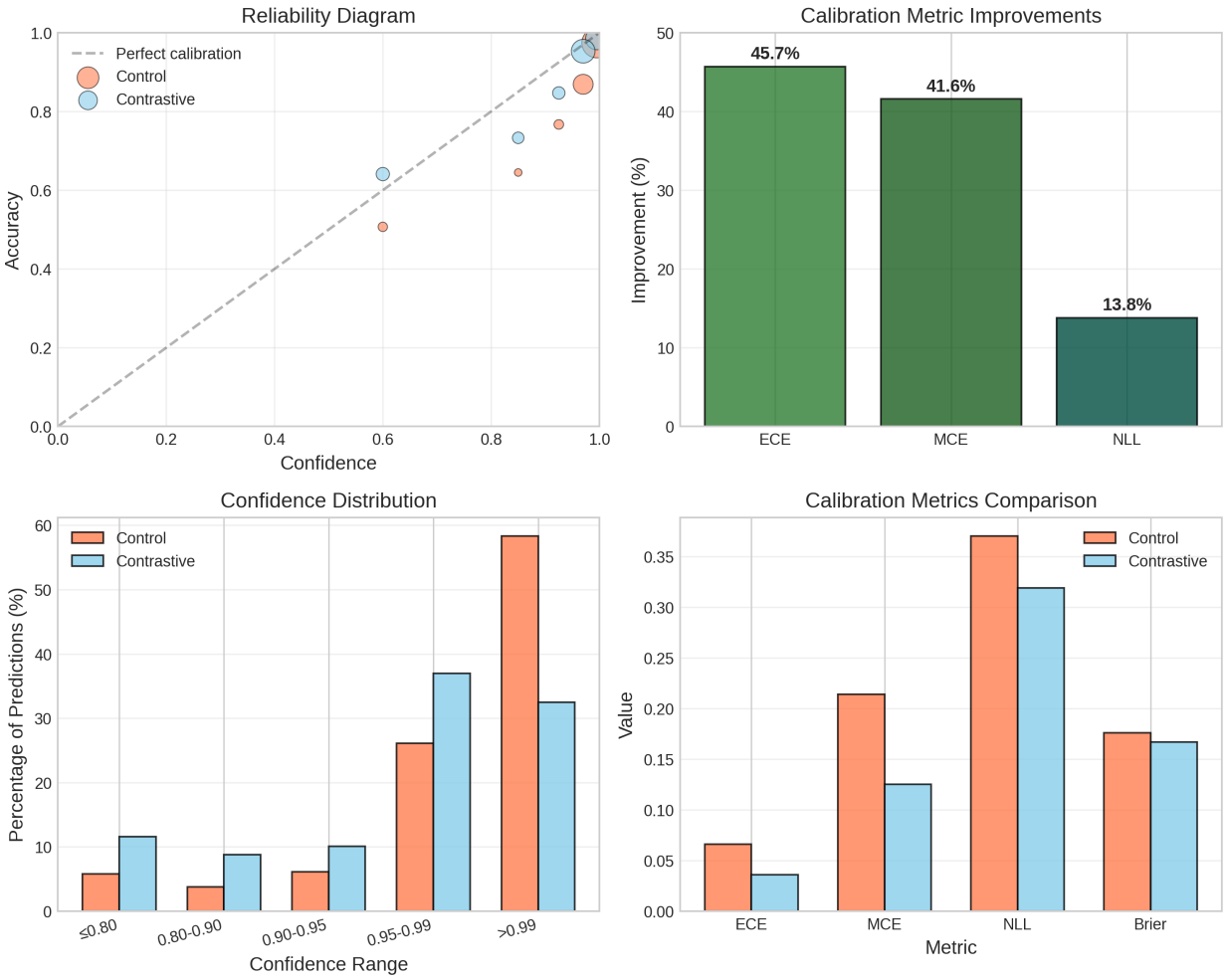

Figure 2: Calibration metrics comparison between control and contrastive models

SNLI Validation Set 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)

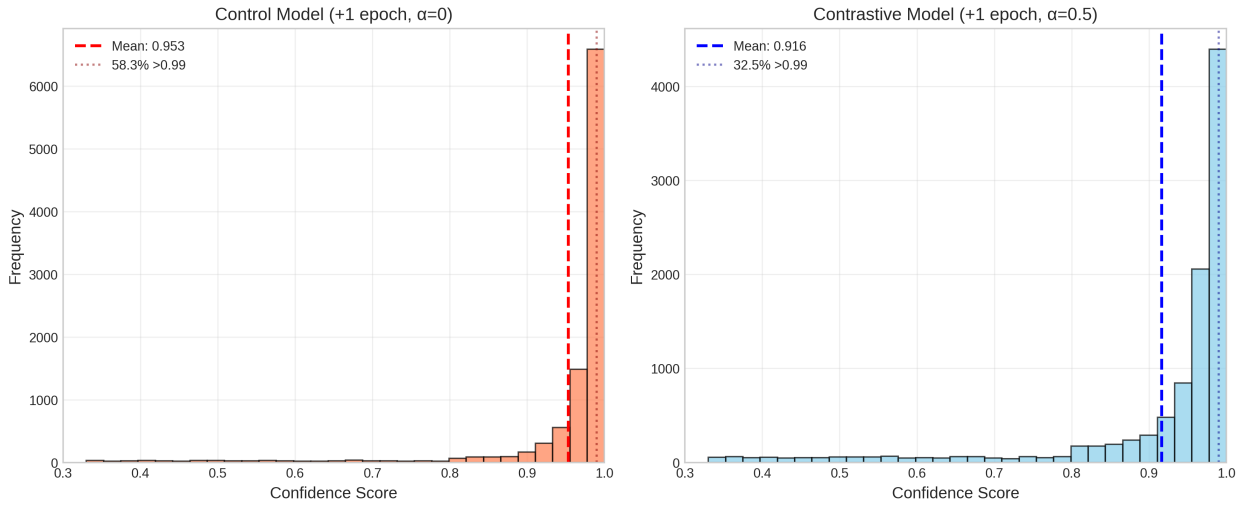

Figure 3: Confidence distribution comparison between control and contrastive models

Confidence Distribution 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.

5.2 Error Analysis

On the challenging ANLI benchmark, contrastive learning shows small but consistent improvements on rounds R2 (+0.7%) and R3 (+1.0%), with no change on R1.

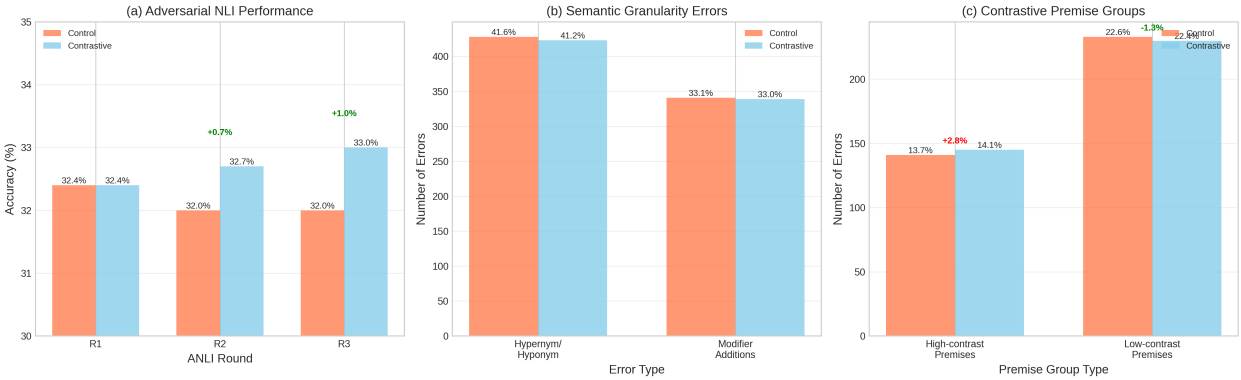

Figure 4: Error Analysis

6. Contrastive 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 reveals the mechanism: the control model places 58.3% of predictions

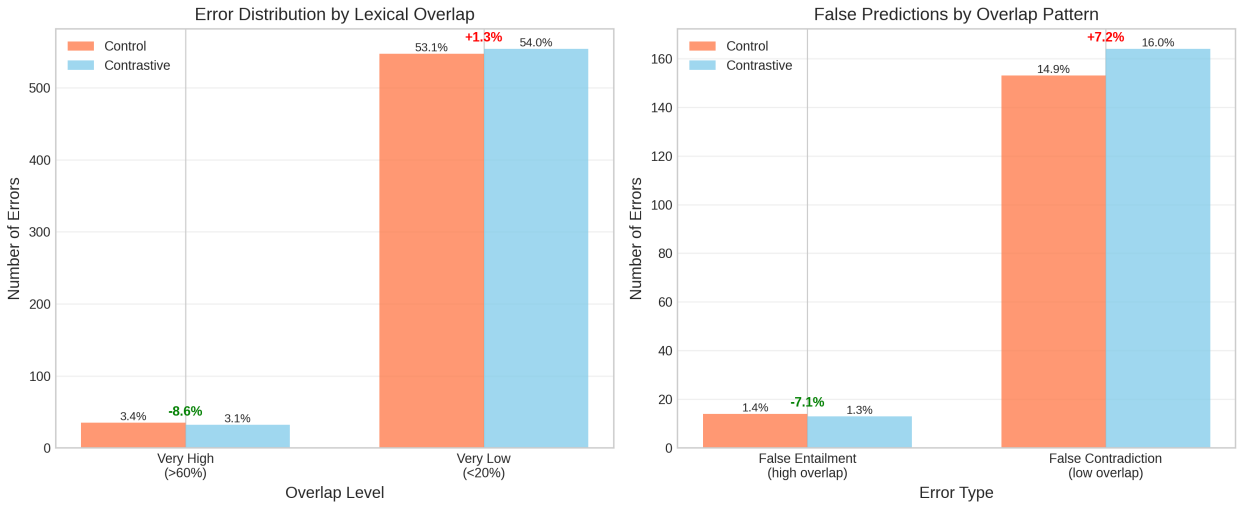

Figure 5: Overlap-based error analysis

at extreme confidence (>0.99), while the contrastive model reduces this to 32.5%. This redistribution doesn't 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.

6.2 Mixed Results on Semantic Categories

Our semantic error analysis shows mixed results across different categories. 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%) 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.

Part II: Adversarial Fine-tuning for Robustness

7. Adversarial Methodology

7.1 Approach

To improve robustness against adversarial examples, we investigate a simple fine-tuning regime that exposes the model to adversarially constructed ANLI examples alongside standard SNLI data. The goal is to "correct" the failure modes exposed by ANLI while maintaining strong SNLI performance.

7.2 Training Strategy

We adopt a joint training approach where SNLI and ANLI datasets are concatenated, allowing the model to learn from both standard and adversarial examples simultaneously. This differs from curriculum-based approaches that might gradually introduce adversarial examples or alternate between datasets.

8. Adversarial Experimental Setup

8.1 Datasets

- **SNLI**: Standard training, validation, and test sets (filtered to remove unlabeled examples)
- **ANLI**: Concatenated train_r1/r2/r3 for training and dev_r1/r2/r3 for validation

ANLI dataset composition:

Round	Context Source	Train Size	Test Size
A1	Wikipedia passages	16,946	1,000
A2	Non-overlapping Wikipedia passages	45,460	1,000
A3	Wikipedia + News + Fiction + Spoken + WikiHow	120,379	1,400

8.2 Training Procedure

1. **Baseline (SNLI-only)**: ELECTRA-small trained 3 epochs on SNLI
2. **Control (SNLI→SNLI)**: Starting from baseline checkpoint, fine-tune 1 additional epoch on SNLI only
3. **Adversarial (SNLI→SNLI+ANLI)**: Starting from baseline checkpoint, fine-tune 1 epoch on concatenated SNLI+ANLI

9. Adversarial Results

9.1 Performance Comparison

Model (Regime)	SNLI	ANLI A1	ANLI A2	ANLI A3	ANLI Combined
Baseline (SNLI-only)	89.29%	30.70%	30.40%	30.83%	30.64%
Control (SNLI→SNLI)	89.53%	31.40%	31.60%	31.83%	31.61%
Adversarial (SNLI→SNLI+ANLI)	89.39%	47.40%	39.50%	40.25%	42.38%

9.2 In-domain (SNLI) Performance

All three models perform nearly identically on SNLI. The control model increases SNLI accuracy by only +0.24 points over baseline, indicating the baseline is near its ceiling. The adversarially fine-tuned model, despite being trained on harder ANLI data, maintains virtually the same SNLI accuracy (-0.14 vs control). This demonstrates that adversarial fine-tuning does not harm in-domain performance.

9.3 Out-of-domain (ANLI) Performance

The baseline model collapses on ANLI (~30.6%, essentially random for 3-way classification). The control model shows minimal improvement (+0.97 points). In contrast, adversarial fine-tuning (SNLI+ANLI) improves dramatically:

- A1: +16.0 points (51.0% relative improvement)
- A2: +7.9 points (25.0% relative improvement)
- A3: +8.4 points (26.4% relative improvement)
- Combined: +10.8 points (34.2% relative improvement)

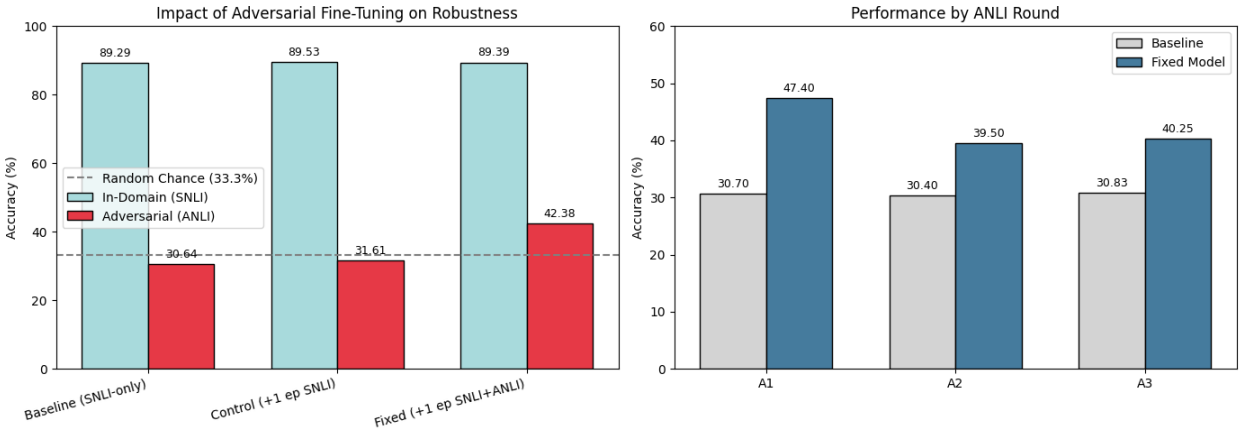

Figure 6: SNLI vs ANLI accuracy across models

9.4 Qualitative Error Analysis

Adversarial fine-tuning fixes several error types that persist in the control model:

Negation trap:

- Premise: "World Premiere ... released by Paramount Pictures"
- Hypothesis: "World Premiere was not released by Universal Studios"
- Gold label: Entailment
- Control prediction: X (Contradiction)
- Adversarial prediction: ✓ (Entailment)

World knowledge:

- Premise: "The Second Jungle Book ... not based on 'The Second Jungle Book' "
- Hypothesis: "The first Jungle Book was written before 1997"
- Gold label: Entailment
- Control prediction: X (Contradiction)
- Adversarial prediction: ✓ (Entailment)

Lexical overlap:

- Premise: "Beat TV ... daily entertainment show with various celebrity guests"
- Hypothesis: "The hosts on Beat TV shared the screen time equally"
- Gold label: Neutral
- Control prediction: X (Entailment)
- Adversarial prediction: ✓ (Neutral)

The adversarially fine-tuned model shows improved handling of negation, better world knowledge integration, and reduced reliance on superficial lexical overlap.

10. Adversarial Discussion

10.1 Robustness Without Trade-offs

The most notable finding is that adversarial fine-tuning substantially improves ANLI performance (+10.8 points combined) without sacrificing SNLI accuracy. This challenges the common assumption that robustness to adversarial examples necessarily comes at the cost of in-domain performance.

10.2 Differential Improvements Across Rounds

The gains are largest on A1 (+16 points) compared to A2 and A3 (+7.9 and +8.4 points respectively). This may reflect:

- A1's simpler domain (Wikipedia only) being more learnable in a single epoch
- Diminishing returns as rounds become progressively harder
- Limited model capacity (ELECTRA-small) struggling with the diverse domains in A3

10.3 Addressing Specific Failure Modes

The qualitative analysis reveals that adversarial fine-tuning helps the model overcome specific shortcuts:

- **Negation handling**: Learning that "not X" doesn't automatically imply contradiction
- **World knowledge**: Better integration of factual information
- **Lexical overlap**: Reduced reliance on surface similarity as a signal for entailment

---

11. Conclusion

We studied two complementary fine-tuning strategies for improving the reliability of an ELECTRA-small NLI model trained on SNLI: premise-bundled contrastive fine-tuning for calibration and adversarial fine-tuning with ANLI for robustness.

On the calibration side, we introduced a contrastive objective that operates over SNLI premise-hypothesis bundles, adding a margin ranking loss between hypotheses with different labels and weighting these pairs by lexical overlap. A single epoch of contrastive fine-tuning left SNLI accuracy essentially unchanged (89.53% $\rightarrow$ 89.58%) but dramatically improved calibration: ECE and MCE dropped by 45.7% and 41.6% respectively, NLL and Brier scores improved, and the fraction of extremely confident predictions (>0.99) fell from 58.3% to 32.5%. Crucially, the proportion of errors made with extreme confidence decreased from 14.5% to 4.7%, indicating that the model learned to express appropriate uncertainty when it was likely to be wrong. These gains came at negligible computational cost.

On the robustness side, we established a strong SNLI baseline and compared two continuation regimes: an additional epoch on SNLI alone (control) and an adversarial regime training on concatenated SNLI+ANLI. While all models maintained similar SNLI performance (~89.3-89.5%), the adversarial run improved ANLI performance dramatically: combined ANLI accuracy increased by +10.8 absolute points (~34% relative) over the control model, with large gains across all ANLI rounds. This shows that adversarial fine-tuning can substantially boost OOD robustness without sacrificing in-domain performance, even with a short additional schedule.

Our findings demonstrate that improving calibration and enhancing robustness require different interventions. The contrastive approach primarily adjusts how models express certainty without changing their decision boundaries, whereas adversarial training teaches models to handle specific failure cases they previously misclassified. While neither strategy alone solves all reliability issues, they provide lightweight, complementary tools for making NLI models more dependable in practice. The dramatic calibration gains achieved with just one additional epoch of contrastive training suggest this approach could be valuable as a lightweight calibration-aware fine-tuning step after standard training.

---

12. Limitations and Future Work

While our fine-tuning strategies deliver strong improvements in calibration and adversarial robustness, there are several opportunities for refinement. First, both contrastive and ANLI-based stages rely on only one additional epoch and a single experimental seed, leaving open questions about stability and optimal training duration. We also restrict our study to the 14M-parameter ELECTRA-small model; larger architectures

may behave differently. Moreover, we primarily evaluate on SNLI and ANLI, so improvements on broader artifact-focused benchmarks (e.g., HANS, contrast sets, hypothesis-only tests) remain to be demonstrated. Future work should therefore examine multi-seed variance, conduct hyperparameter sweeps over contrastive margins and dataset mixing ratios, and explore extended or interleaved training schedules. Enhancing the quality and diversity of contrastive examples—through structured or synthetic generation targeting specific linguistic phenomena—may further improve calibration. Finally, joint optimization of calibration and robustness objectives, along with scaling to larger models and multi-genre NLI datasets, could validate the generality of these approaches and strengthen claims of reduced shortcut reliance.

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*.

Nie, Y., et al. (2020). Adversarial NLI: A new benchmark for natural language understanding.

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*.