Abstract

Natural language inference (NLI) models based on pre-trained transformers can achieve high accuracy on benchmarks like SNLI, yet they often exhibit two intertwined failures: overconfident probabilities and brittle behavior on adversarial challenge sets. In this work we study two complementary fine-tuning strategies for ELECTRA-small: (1) a calibration-oriented contrastive learning objective that exploits naturally occurring premise-hypothesis bundles in SNLI, and (2) an adversarial robustness-oriented fine-tuning stage that mixes ANLI with SNLI. Our contrastive approach adds a margin ranking loss with lexical-overlap-based dynamic weighting on hypothesis pairs that share a premise but have different labels, encouraging the model to express uncertainty on fine-grained lexical and semantic distinctions. On SNLI, this contrastive fine-tuning leaves accuracy essentially unchanged (89.53% $\rightarrow$ 89.58%) but dramatically improves calibration: Expected Calibration Error drops by 45.7% (0.066 $\rightarrow$ 0.036), Maximum Calibration Error by 41.6%, Negative Log-Likelihood by 13.8%, and the fraction of overconfident predictions (>0.99) from 58.3% to 32.5%, with extreme high-confidence errors reduced from 14.5% to 4.7% of all errors. On ANLI, it yields small but consistent accuracy gains on harder rounds (+0.7 on R2, +1.0 on R3). In a complementary setup, we perform a brief adversarial fine-tuning phase on concatenated SNLI+ANLI. Relative to a control model trained for an extra epoch on SNLI alone, this adversarial regime improves combined ANLI accuracy by +10.8 absolute points (~34% relative) while maintaining SNLI accuracy (~89.4%), demonstrating substantial robustness gains without sacrificing in-domain performance. Taken together, our results show that calibration-aware contrastive fine-tuning and ANLI-based adversarial fine-tuning address distinct but synergistic aspects of NLI reliability: the former reshapes confidence distributions with minimal cost, while the latter substantially improves performance on challenging out-of-distribution data.

---

1. Introduction

Natural language inference (NLI) is a core benchmark for evaluating language understanding, requiring models to predict whether a hypothesis is entailed by, contradicts, or is neutral with respect to a premise (Bowman et al., 2015). Transformer-based models fine-tuned on SNLI routinely achieve high accuracy, but a growing body of work shows that this success often relies on superficial cues and dataset artifacts rather than robust reasoning. Hypothesis-only baselines, lexical overlap heuristics, and simple negation cues can explain much of the performance (Poliak et al., 2018; McCoy et al., 2019; Gardner et al., 2020), raising concerns about how well these models generalize beyond benchmark test sets.

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 such as ANLI (Nie et al., 2020) are constructed in a human-and-model-in-the-loop manner to expose model weaknesses through subtle perturbations (lexical substitutions, negation, world knowledge). Models that perform well on SNLI often collapse to near-random accuracy 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

A central critique of NLI benchmarks is that they contain annotation artifacts—systematic correlations between surface patterns and labels that models can exploit without genuine reasoning. Poliak et al. (2018) show that hypothesis-only models achieve surprisingly strong performance on multiple NLI datasets, indicating that premise information is often unnecessary. McCoy et al. (2019) identify syntactic heuristics (e.g., "high lexical overlap implies entailment") that models mistakenly internalize. Gardner et al. (2020) introduce contrast sets—minimally edited examples that flip the gold label—to demonstrate how brittle these shortcuts are under small perturbations.

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

A probabilistic classifier is well-calibrated when its confidence scores match empirical correctness frequencies (Guo et al., 2017). Modern deep networks tend to be overconfident, especially when trained with cross-entropy and aggressive regularization or data augmentation. Poor calibration is particularly problematic when outputs are consumed by downstream decision-making components, thresholding rules, or risk-sensitive systems.

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) is a pre-trained transformer that replaces masked language modeling with a discriminative pre-training task: distinguishing real tokens from synthetic replacements. ELECTRA-small provides a strong, efficient encoder backbone for classification tasks like NLI; we use it as our shared base model in all experiments.

To assess genuine language understanding and robustness, several adversarial challenge datasets have been introduced. Among them, ANLI (Nie et al., 2020) is notable for its human-and-model-in-the-loop construction: annotators iteratively craft adversarial NLI examples that fool a current model. ANLI is released in three rounds (A1-A3) with different domains and increasing difficulty, including Wikipedia, news, fiction, spoken language, and instructional texts. Models trained only on SNLI often achieve near-random accuracy (~33%) on ANLI, reflecting severe out-of-distribution (OOD) brittleness.

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 been highly successful in representation learning (Chen et al., 2020) and has recently been adapted to NLP tasks. In reading comprehension, Dua et al. (2021) introduce instance bundles—groups of examples sharing a context but with different answers—and apply contrastive objectives to encourage robust distinctions, showing improvements on adversarial evaluation sets.

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.

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).

4.3 Hyperparameters

- Batch size: 32
- Learning rate: 5e-5 with linear warmup
- Optimizer: AdamW
- Max sequence length: 128 tokens

4.4 Evaluation Metrics

We evaluate on: - Overall accuracy on SNLI and ANLI - Confidence calibration metrics (ECE, MCE, NLL, Brier score) - Fine-grained error categorization based on linguistic phenomena - Performance on high vs. low lexical overlap subsets

5. Contrastive Results

5.1 Overall Performance and Calibration

SNLI Validation Set

Model	Accuracy	ECE	MCE	NLL	Brier
Control (+1 epoch, α=0)	89.53%	0.066	0.214	0.370	0.176
Contrastive (+1 epoch, α=0.5)	89.58%	0.036	0.125	0.319	0.167

The calibration metrics reveal substantial improvements despite minimal accuracy gains (+0.05%):

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

Confidence Distribution

Confidence Range	Control		Contrastive	
	Count (%)	Accuracy	Count (%)	Accuracy
(0.99, 1.00]	58.3%	0.974	32.5%	0.985
(0.95, 0.99]	26.1%	0.869	37.0%	0.953
(0.90, 0.95]	6.1%	0.767	10.1%	0.847
(0.80, 0.90]	3.8%	0.645	8.8%	0.733
$\leq$ 0.80	5.8%	0.507	11.6%	0.641

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.

Adversarial NLI (Out-of-Distribution)

Model	ANLI-R1	ANLI-R2	ANLI-R3
Control	32.4%	32.0%	32.0%
Contrastive	32.4%	32.7%	33.0%

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.

5.2 Error Analysis by Category

Primary Targets (Lexical Contrasts)

Category	Control Errors	Contrastive Errors	Change
Negation flips	18 (1.7%)	19 (1.9%)	+5.6%
Quantifier changes	164 (15.9%)	174 (17.0%)	+6.1%
Antonym substitutions	57 (5.5%)	59 (5.8%)	+3.5%

Secondary Targets (Semantic Granularity)

Category	Control Errors	Contrastive Errors	Change
Hypernym/hyponym	428 (41.6%)	423 (41.2%)	-1.2%
Modifier additions	341 (33.1%)	339 (33.0%)	-0.6%

Contrastive Premise Groups

Category	Control Errors	Contrastive Errors	Change
High-contrast premises	141 (13.7%)	145 (14.1%)	+2.8%
Low-contrast premises	233 (22.6%)	230 (22.4%)	-1.3%

5.3 Overlap-Based Analysis

Overlap Level	Control Errors	Contrastive Errors	Change
Very high (>60%)	35 (3.4%)	32 (3.1%)	-8.6%
Very low (<20%)	547 (53.1%)	554 (54.0%)	+1.3%
False entailment (high overlap)	14 (1.4%)	13 (1.3%)	-7.1%
False contradiction (low overlap)	153 (14.9%)	164 (16.0%)	+7.2%

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 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 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. 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

8.3 Hyperparameters

- Batch size: 8
- Learning rate: 5e-5 with linear scheduler, no warmup
- Optimizer: AdamW (β_1=0.9, β_2=0.999, ϵ=1e-8)
- Max sequence length: 128 tokens
- Weight decay: 0.0, max grad norm: 1.0

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)

9.4 Qualitative Error Analysis

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

Error Type	Example	Control	Adversarial
Negation trap	P: “World Premiere … released by Paramount Pictures”H: “World Premiere was not released by Universal Studios”Gold: Entailment	X (Contradiction)	✓ (Entailment)
World knowledge	P: “The Second Jungle Book … not based on ‘The Second Jungle Book’ ”H: “The first Jungle Book was written before 1997”Gold: Entailment	X (Contradiction)	✓ (Entailment)
Lexical overlap	P: “Beat TV … daily entertainment show with various celebrity guests”H: “The hosts on Beat TV shared the screen time equally”Gold: Neutral	X (Entailment)	✓ (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.

Taken together, our results highlight that calibration and robustness are related but distinct objectives. Contrastive fine-tuning mainly reshapes the model's confidence distribution, making its probabilities more trustworthy, while adversarial fine-tuning mainly improves correctness on challenging adversarial examples. 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

Our work has several limitations that should be addressed in future research:

1. **Limited training duration**: Both contrastive and adversarial fine-tuning use only one additional epoch, which may be insufficient for optimal performance.
2. **Single seed experiments**: All results are from single runs without variance estimates across random seeds.
3. **Model size constraints**: We use ELECTRA-small (14M parameters) for efficiency, but larger models might show different patterns.
4. **Limited adversarial evaluation**: While we evaluate on ANLI, more comprehensive testing on HANS, contrast sets, and hypothesis-only baselines would strengthen claims about artifact mitigation.
5. **Static hyperparameters**: We use fixed margins and weighting schemes without extensive tuning.
6. **Bundle quality**: Natural premise-hypothesis bundles in SNLI may not consistently contain the contrasts needed for specific phenomena.

---

13. Future Work

Our findings suggest several directions for extending and combining calibration- and robustness-oriented fine-tuning:

1. **Direct artifact and challenge-set testing.** Evaluate both contrastive and adversarially fine-tuned models on targeted artifact benchmarks such

as HANS and contrast sets, as well as hypothesis-only tests, to more directly quantify reductions in shortcut reliance.

2. **Multiple seeds and hyperparameter sweeps.** Explore variance across random seeds and investigate the sensitivity of both methods to the contrastive margin, overlap weighting function, mixing ratio of SNLI/ANLI, and number of fine-tuning epochs.

3. **Extended and interleaved training schedules.** Study longer schedules, interleaving or alternating contrastive and adversarial batches, and curriculum strategies that gradually increase difficulty.

4. **Synthetic and structured contrast generation.** Generate additional synthetic contrast pairs targeting specific linguistic phenomena such as negation, quantifiers, and world knowledge, then integrate them into the contrastive objective.

5. **Fine-grained calibration analysis.** Stratify calibration by overlap level, label type, and linguistic category, investigating whether high-overlap entailment decisions become better calibrated than low-overlap contradictions.

6. **Combining calibration and robustness objectives.** Jointly optimize for calibration and robustness by applying overlap-weighted contrastive losses on both SNLI and ANLI examples or adding calibration-aware regularizers during adversarial fine-tuning.

7. **Scaling to larger models and other datasets.** Test whether our findings hold for larger ELECTRA variants or other architectures, and whether similar benefits arise on multi-genre NLI datasets such as MultiNLI.

8. **Curriculum learning strategies.** Explore gradual mixing of SNLI and ANLI data, starting with easier adversarial examples and progressively increasing difficulty.

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*.