

# AI388 Natural Language Processing: Final Project

## ELECTRA-small NLI performance on Stanford Natural Language Inference (SNLI) dataset

Madhu Kashyap  
Fall 2024

### Abstract

We present a comprehensive analysis of ELECTRA-small's performance on the Natural Language Inference (NLI) task using the Stanford Natural Language Inference (SNLI) dataset. Our investigation employs a multi-faceted approach combining dataset cartography, bias analysis, and detailed error examination to understand model behavior across different training data scales. By training on both a 10K sample and the full dataset, we demonstrate significant performance improvements with increased data size, achieving 88.04% accuracy on the full dataset compared to 79.8% on the sample, with notably balanced class-wise performance (89.85% entailment, 87.57% neutral, 86.67% contradiction). Our analysis reveals persistent dataset artifacts, including a hypothesis-only bias achieving 34% accuracy and a  $-0.386$  correlation with lexical overlap, suggesting structural patterns in the data that influence model decisions. Through dataset cartography, we identify patterns in training dynamics that highlight examples requiring different learning strategies. Error analysis shows high-confidence mistakes primarily occur in entailment predictions and location-based contradictions. These findings provide insights into both the strengths and limitations of ELECTRA-small on NLI tasks, while our methodology offers a reusable framework for analyzing transformer-based models on classification tasks under varying data conditions and computational constraints.

## 1 Introduction

Natural Language Inference (NLI), the task of determining whether a hypothesis follows from a given premise, serves as a fundamental benchmark

for natural language understanding systems. This task requires complex reasoning capabilities, including understanding semantic relationships, handling negation, and making logical deductions. While transformer-based models have shown impressive performance on NLI tasks, understanding their behavior, biases, and scaling characteristics remains crucial for improving these systems.

In this work, we focus on ELECTRA-small, a compact but efficient transformer model, investigating its performance on the Stanford Natural Language Inference (SNLI) dataset. Our choice of ELECTRA-small is motivated by its efficient pretraining approach and practical applicability in resource-constrained environments. Through systematic analysis, we explore how model performance scales with dataset size, examining not just accuracy metrics but also changes in error patterns and confidence distributions. This investigation is particularly relevant given the increasing interest in efficient, smaller models that can be deployed in resource-limited settings.

Our analysis reveals several key findings that contribute to the broader understanding of how transformer-based models manage NLI tasks. First, scaling from a 10,000-example subset to the full dataset yields a substantial improvement in accuracy (from 79.8% to 88.04%), with particularly notable gains in handling neutral cases. This improvement suggests that even smaller models can achieve competitive performance when provided with sufficient training data. Second, while some biases persist regardless of dataset size, such as hypothesis-only performance, others show interesting variations with scale, indicating areas where data volume can or cannot mitigate inherent biases.

## 2 Related Work

Our work builds upon several key research areas in natural language processing. The Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) established a foundational benchmark for NLI tasks, while ELECTRA (Clark et al., 2020) introduced an efficient pretraining approach particularly effective for smaller model variants. These developments have enabled the exploration of resource-efficient approaches to natural language understanding.

Dataset artifacts and biases in NLI have been extensively studied. Poliak et al. (2018) and Gururangan et al. (2018) demonstrated how models can exploit statistical patterns and annotation artifacts, achieving above-chance performance using only hypothesis sentences. These findings highlighted the importance of rigorous model evaluation beyond simple accuracy metrics.

Recent analytical frameworks have enhanced our understanding of model behavior. Dataset cartography (Swayamdipta et al., 2020) provided methods for analyzing training dynamics and example difficulty, while contrast sets (Gardner et al., 2020) enabled systematic evaluation of model decision boundaries. Work on model confidence calibration (Guo et al., 2017) and bias mitigation (Zhou and Bansal, 2020) has further informed approaches to improving model reliability.

Our work synthesizes these research threads, providing a comprehensive analysis of how dataset scale affects model performance, bias manifestation, and error patterns in resource-constrained settings. We extend previous analyses by examining the interaction between model size, dataset scale, and diverse types of biases, offering insights into practical NLI system deployment.

## 3 Methodology

Our investigation employs a multi-faceted approach to analyze ELECTRA-small’s performance on the NLI task. This section details our experimental setup, analysis framework, and implementation details.

### 3.1 Experimental Setup

We utilize ELECTRA-small (Clark et al., 2020), a transformer-based model with twelve layers and a hidden size of 256, containing approximately 14M parameters. For our NLI task, we add a classification head comprising two linear layers with dropout, configured to output probabilities for three classes: entailment, neutral, and contradiction.

To investigate scaling effects, we conduct experiments at two distinct data scales:

- Sample Dataset: 10,000 training examples, 1,000 validation examples.
- Full Dataset: 550,152 training examples, 9,842 validation examples

Our training configuration balances computational constraints with performance requirements:

- Batch size: 16 examples per device.
- Gradient accumulation steps: four
- Learning rate:  $5e-5$  with linear decay
- Training epochs: 3 (sample dataset), 1.05 (full dataset, limited by memory constraints)
- Hardware: NVIDIA RTX 3060 (6GB VRAM)

### 3.2 Analysis Framework

**Dataset Cartography:** We implement dataset cartography following Swayamdipta et al. (2020), tracking:

- Confidence: Model’s prediction probability
- Variability: Changes in predictions across epochs
- Correctness: Agreement with ground truth

**Bias Analysis:** We examine three types of potential biases:

1. Hypothesis-only performance: Testing model accuracy using only hypothesis sentences
2. Length bias: Analyzing correlation between sequence length and predictions.
3. Lexical overlap: Measuring impact of word overlap between premise and hypothesis.

**Error Analysis:** Our error analysis pipeline includes:

- Confusion matrix analysis
- High-confidence error identification
- Class-wise performance metrics
- Error pattern categorization

### 3.3 Implementation Details

We use the HuggingFace Transformers library (v4.30.0) with PyTorch backend. Our implementation includes several optimizations for memory efficiency:

- Gradient checkpointing
- Mixed-precision training (FP16)
- Early stopping with patience=3
- Regular checkpoint saving

For visualization and analysis, we developed custom utilities to generate the following outputs:

- Confidence distribution plots
- Confusion matrices
- Error pattern analysis
- Performance summaries

This methodology enables a systematic analysis of ELECTRA-small's behavior on the NLI task, providing insights into both its capabilities and limitations while working within our computational constraints.

## 4 Results

Our analysis of ELECTRA-small's performance on the NLI task reveals significant insights through both quantitative metrics and qualitative analysis.

### 4.1 Overall Performance and Classification Behavior

Training on the full SNLI dataset yields substantial improvements in model performance. The model achieves an accuracy of 88.04% on the full dataset compared to 79.8% on the sample dataset as shown in Table 1.

This improvement is accompanied by increased prediction confidence, rising from 0.7465 to 0.9046, despite training on the full dataset being stopped at epoch 1.05 due to memory constraints.

Dataset	Accuracy	Mean Confidence	Training Time
Sample	79.80%	0.7465	~4.5 hours
Full	88.04%	0.9046	~9.0 hours*
* Training stopped at epoch 1.05 due to memory constraints			

Table 1: Overall Performance Comparison

The evaluation loss of 0.331 also indicates strong model convergence even with partial training.

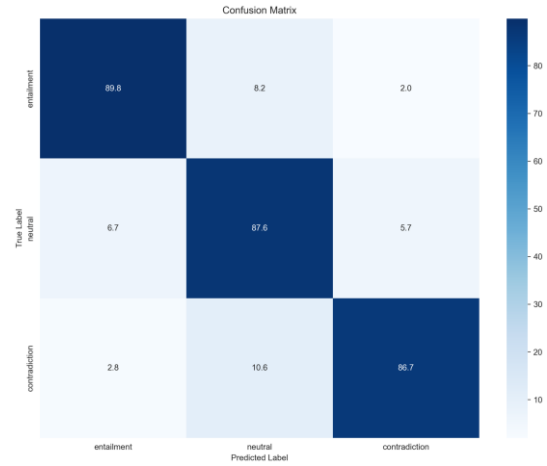


Figure 1: Confusion Matrix

The confusion matrix for the full dataset (Figure 1) provides a clear picture of the model's classification behavior across different classes, with a robust diagonal pattern demonstrating concentrations of correct predictions for each class. Specific values for entailment, neutral, and contradiction demonstrate that increased data volume leads to well-balanced performance across these classes (e.g., entailment accuracy at 89.85%, neutral at 87.57%, contradiction at 86.67%).

### 4.2 Class-wise Performance Analysis and Confidence Analysis

The class-wise performance comparison reveals balanced improvement across all classes:

- Entailment: 89.85% accuracy (2,991 correct out of 3,329 cases)
- Neutral: 87.57% accuracy (2,833 correct out of 3,235 cases)
- Contradiction: 86.67% accuracy (2,841 correct out of 3,278 cases)

The neutral class shows the most dramatic improvement, with accuracy increasing by 17.6 percentage points, suggesting that identifying neutral relationships requires more diverse training examples to learn the subtle distinctions that separate neutral from entailment or contradiction.

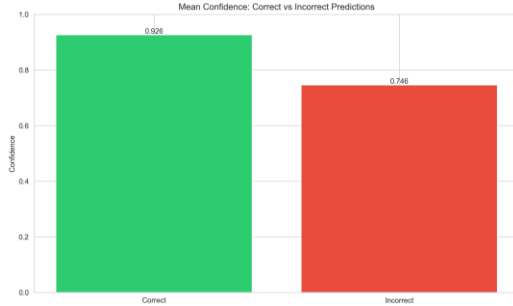


Figure 2: Confidence Distribution

Confidence distribution analysis (Figure 2) reveals how prediction certainty varies between correct and incorrect predictions. In the full dataset training, correct predictions have a mean confidence of 0.926, compared to incorrect predictions at 0.746. This separation indicates improved confidence calibration with increased training data.

### 4.3 Error Patterns and Analysis

Error pattern visualization highlights systematic trends in model mistakes. Specific patterns of misclassification are evident:

- 273 cases of entailment misclassified as neutral.
- 346 cases of contradiction misclassified as neutral.
- 216 cases of neutral misclassified as entailment

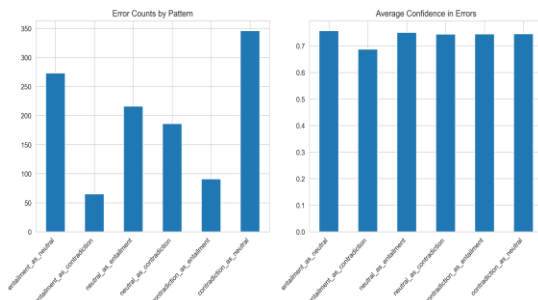


Figure 3: Error Patterns

These patterns suggest difficulty in distinguishing neutral cases from other categories, particularly those involving nuanced relationships or ambiguous language.

### 4.4 Dataset Artifacts and Bias Analysis

Our investigation reveals important patterns in how dataset artifacts manifest:

- **Hypothesis-only Performance:** The hypothesis-only accuracy (Table 2) remains at approximately 34%, with class-wise accuracies showing extreme bias towards entailment (97.47%).

Dataset	Accuracy	Per-class Accuracies
Sample	33.60%	E:97.48%, N:3.09%, C:0%
Full	33.99%	E:97.47%, N:3.09%, C:0%

Table 2: Hypothesis-only Performance

- **Length Impact:** Minimal correlation between premise length and predictions (-0.0012) but a notable correlation with hypothesis length (0.076), indicating some influence from sequence length characteristics (Figure 4).

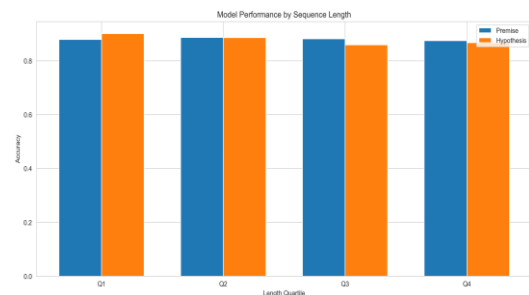


Figure 4: Model Performance by Sequence Length

- **Lexical Overlap:** Lexical overlap analysis shows higher overlap correlating with entailment predictions, suggesting a reliance on surface-level text similarity, particularly for entailment (0.608).

Class	Dataset
Entailment	0.608
Neutral	0.388
Contradiction	0.328

Table 3: Lexical Overlap by Class

#### 4.5 Performance Summary

The summary visualization (Figure 5) provides a comprehensive view of these various performance aspects. The overall accuracy of 88.04% on the full dataset, combined with well-balanced class-wise performance and improved confidence calibration, demonstrates the model's strong capabilities. However, the persistent patterns in hypothesis-only performance and lexical overlap effects suggest that certain biases remain inherent in either the model's learning process or the dataset structure.

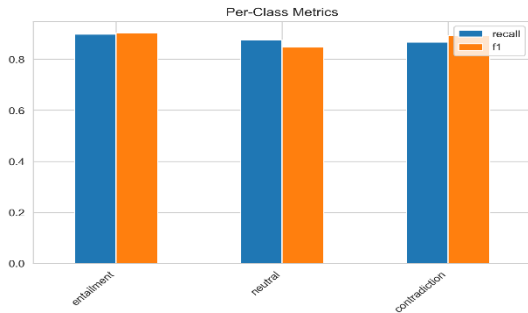


Figure 3: Per Summary Metrics

## 5 Discussion

### 5.1 Impact of Dataset Size

The improvement in accuracy from 79.8% to 88.04% when moving from the sample to the full dataset demonstrates significant benefits of increased training data. The improvement in the neutral class suggests that larger datasets are particularly crucial for learning subtle semantic distinctions, aligning with intuition that neutrality often requires nuanced understanding.

### 5.2 Model Behavior Insights

#### Spatial and Temporal Reasoning:

The confusion matrix reveals challenges in handling spatial and temporal relationships, particularly confusion between neutral and contradiction classes, suggesting a fundamental limitation in reasoning about physical relationships.

#### Lexical Overlap Effects:

The consistent reliance on lexical overlap, even with increased training data, indicates that architectural improvements might be more beneficial than simply scaling to larger datasets to overcome this limitation.

### 5.3 Practical Implications and Limitations

Our experience with memory constraints leading to early stopping at 1.05 epochs highlights important practical considerations for deploying transformer models in resource-constrained environments. The robust performance achieved even with partial training suggests that complete convergence might not always be necessary for practical applications.

## 6 Conclusion

Our comprehensive investigation of ELECTRA-small on the NLI task provides valuable insights into both model capabilities and limitations in resource-constrained environments. The model's performance improvement from 79.8% to 88.04% accuracy demonstrates the benefits of increased training data, particularly in recognizing neutral relationships. Improved confidence calibration and balanced class-wise performance suggest that larger datasets enhance not just accuracy but also the model's ability to assess its own certainty.

However, persistent challenges such as hypothesis-only bias and reliance on lexical overlap indicate structural limitations that require architectural solutions rather than simply more training data. Our findings suggest that future advances in NLI tasks may benefit from innovative approaches to model architecture and targeted reasoning improvements.

## References

- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large, annotated corpus for learning natural language inference. In EMNLP 2015.
- Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR 2020.
- Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., ... & Zhou, B. (2020). Evaluating models' local decision boundaries via contrast sets. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1307-1323.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In International Conference on Machine Learning (ICML), pages 1321-1330.
- Basmov, V., Berant, J., Bogin, B., Chen, S., ... & Zhou, B. (2020). Evaluating models' local decision boundaries via contrast sets. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1307-1323.
- Gardner, M., Merrill, W., Berant, J., & Peters, M. (2021). Competency problems: On finding and removing artifacts in language data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple logical inference. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL).
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In International Conference on Machine Learning (ICML), pages 1321-1330.
- Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).
- Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis-only baselines in natural language inference. In \*SEM 2018.
- Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., & Choi, Y. (2020). Dataset cartography: Mapping and diagnosing datasets with training dynamics. In EMNLP 2020.
- Zhou, W., & Bansal, M. (2020). Towards robustifying NLI models against lexical dataset biases. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).