Enhancing ELECTRA-small with Dataset Cartography: Probing Linguistic Sensitivity and Robustness on SNLI

Soo Ihk Ro

Reese Williamson

Abstract

In this paper, we evaluate the performance of the ELECTRA-small model on the Stanford Natural Language Inference (SNLI) dataset and conduct an in-depth analysis to understand the model's robustness and underlying learning patterns. Specifically, we construct a contrast set for the SNLI dataset by altering only the hypothesis of each sentence using three distinct transformation methods: synonym substitution, syntactic rephrasing, and semantic shift. These modifications allow us to probe the model's sensitivity to linguistic variations and its reliance on shallow heuristics. Our findings reveal notable weaknesses in ELECTRA-small's ability to handle nuanced changes in language, with significant drops in performance across contrast sets. To address these challenges, we implement a retraining strategy that targets the error-prone subsets of the dataset, using Dataset Cartography to identify and focus on ambiguous and challenging examples. The retraining approach demonstrates improvements in model robustness, suggesting that targeted fine-tuning can effectively mitigate model reliance on spurious correlations and enhance its generalization capabilities.

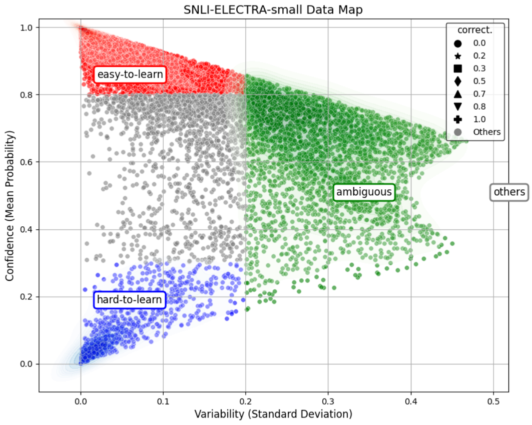


Figure 1: Data Map for SNLI 25,000 train set, based on ELECTRA-small classifier. The Dataset cartography map displays model confidence (mean probability) and variability (standard deviation) categorized as easy-to-learn (top left), ambiguous (right), hard-to-learn (bottom left), and others (left)."

1 Introduction

Natural language processing (NLP) has always faced challenges in accurately modeling language entailment tasks. With the rise of deep learning and transformer models, substantial progress has been made in addressing these challenges. Transformer-based models have achieved state-of-the-art performance in various tasks, such as question answering, language modeling, and text classification. Among these models, ELECTRA has gained attention for its efficiency and performance relative to its size, with its small variant providing an optimal balance between performance and computational efficiency for resource-constrained applications (Clark et al., 2020).

In this study, we fine-tune the ELECTRA-small model on the SNLI dataset, a well-established benchmark for natural language inference. However, it contains annotation artifacts that models often exploit, leading to superficial pattern learning. To address these gaps, we create contrast sets by modifying hypotheses through synonym substitution, syntactic rephrasing, and semantic shifts. These modifications often change the gold label, allowing us to examine how well the model handles deviations from its training distribution. Our analysis reveals that the model relies on shallow patterns, making it vulnerable to subtle linguistic changes. As proposed by Gardner et al. (2020), contrast sets systematically modify validation or test examples to expose weaknesses in model decision boundaries and assess generalization more effectively.

We then implement a targeted retraining strategy using Dataset Cartography (Swayamdipta et al., 2020), focusing on data subsets where the model performs poorly. Dataset Cartography helps visualize and categorize instances based on training dynamics, identifying ambiguous and hard-to-learn examples. By emphasizing these challenging examples during retraining, we improve the model's ability to generalize to nuanced variations and reduce reliance on superficial correlations.

Our contributions are threefold: (1) we evaluate the ELECTRA-small model's performance on systematically modified validation sets, (2) we identify error patterns and analyze their implications for model robustness, and (3) we demonstrate the effectiveness of targeted retraining for handling nuanced linguistic variations. These results highlight the value of contrast set evaluation and focused fine-tuning to improve NLP model resilience.

2 Model, Dataset and Contrast Sets

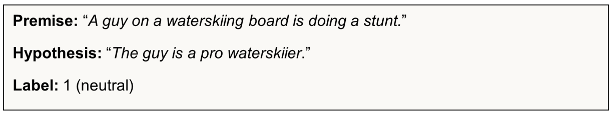
2.1 The Model

For our investigation, the "Efficiently Learning an Encoder that Classifies Token Replacements Accurately" (ELECTRA)-small model was utilized. ELECTRA-small is a lightweight variant model of the ELECTRA architecture designed for balance between performance and computational efficiency. The main parameters of the model consist of 12 transformer layers with a hidden size of 256 and a feed-forward network inner hidden size of 1024. The model uses 4 attention heads with each head having the size of 64 and an embedding size of 128.

Rather than the traditional masked language modeling approach, ELECTRA-small utilizes replaced token detection. This pre-training strategy enables the model to identify whether each token in the input was replaced by a generator model, providing a more sample-efficient training process. The reduced size and efficient training objective make ELECTRA-small an attractive option for resource-constrained environments while still delivering strong performance on various NLP benchmarks, including the Stanford NLI dataset (Clark et al., 2020).

2.2 Stanford Natural Language Inference Dataset

This written work utilizes the Stanford NLI (SNLI) dataset, which consists of 550,152 training examples, along with 10,000 examples each for validation and testing.



Premise: "*A guy on a waterskiing board is doing a stunt.*"

Hypothesis: "*The guy is a pro waterskiier.*"

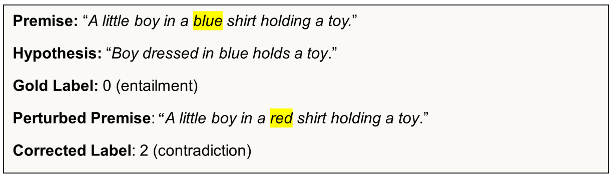
Label: 1 (neutral)

Figure 2: Stanford NLI dataset example

As shown in **Figure 2** each test example includes three main components: a premise, which is a sentence or sentence fragment describing a scene; a hypothesis, which is another sentence or fragment making a claim or describing a scene; and a label, indicating the relationship between the premise and hypothesis. The label can be entailment (hypothesis follows from the premise, labeled as 0), neutral (no clear inference, labeled as 1), or contradiction (hypothesis contradicts the premise, labeled as 2) (Bowman et al., 2015).

2.3 Premise Contrast Set

Our contrast set methodology involved modifying the premise of 600 examples from the SNLI dataset to challenge the model's understanding of overall context rather than relying on superficial cues from the hypothesis. These minor perturbations to the premise often altered its meaning and, in many cases, led to changes in the example's label. By modifying the premise, we provided a more rigorous scenario for detecting shallow heuristics, forcing the model to process new foundational information and reassess the relationship dynamically. This approach reveals weaknesses in the model's ability to comprehend complex entailment relationships and ensures that its understanding of context is both deep and authentic.



Premise: "*A little boy in a blue shirt holding a toy.*"

Hypothesis: "*Boy dressed in blue holds a toy.*"

Gold Label: 0 (entailment)

Perturbed Premise: "*A little boy in a red shirt holding a toy.*"

Corrected Label: 2 (contradiction)

Figure 3: Contrast Set example of Premise Modification

Figure 3 illustrates an example of how we minimally altered premise contrast sets. Such perturbations frequently change the original label, for instance, shifting it from entailment (label 0) or neutral (label 1) to contradiction (label 2). In Figure 3, the original label was entailment (0), as the premise supported the hypothesis. By changing ”blue” to ”red” in the premise, the label shifted to contradiction (2) because the altered premise no longer supports the hypothesis.

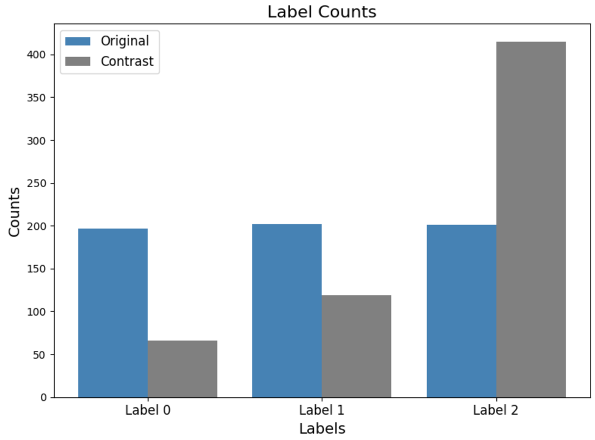


Figure 4: Label distribution, Original Data vs. Premise Contrast Set

However, modifying the premise to transform a contradictory or unrelated hypothesis into a supportive one is often impractical. If the hypothesis is entirely unrelated to the premise, achieving entailment would require substantial changes to the premise, far beyond minor perturbations, thereby violating the principles of contrast sets (Gardner et al., 2020). Consequently, our contrast sets contain a higher proportion of examples labeled as contradiction (2) compared to the original dataset. **Figure 4** illustrates this distribution: while the original SNLI dataset had 199 examples labeled as 0, 190 as 1, and 200 as 2, the modified contrast set had 66 examples labeled as 0, 119 as 1, and 415 as 2.

2.4 Hypothesis Contrast Set

Next we explored the method of minimally perturbing the hypothesis of each example instead of the premise. **Figure 4** shows an imbalance in types of labels for our contrast set, thus we sought to make a new one with a more balanced set. A 600 example hypothesis perturbation contrast set was manually created as the hypothesis itself directly engages with the relationship classification (entailment, contradiction, neutral). Changing just the hypothesis and leaving the premise intact maintains the context for the original task. As altering the premise risks introducing ambiguities with the original task's intent. By ensuring the context remains consistent, this approach isolates the model's ability to infer relationships without the confounding factor of a changing premise, making label adjustments more straightforward (Sanwal, 2024).

We utilized three methods for hypothesis alterations: synonym substitution, syntactic rephrasing, and semantic shifts. Synonym substitution involved replacing key words (e.g., nouns, verbs, adjectives) with their synonyms to assess whether the model could handle vocabulary changes without altering its predictions, thus demonstrating an understanding of meaning beyond specific terms. Syntactic rephrasing focused on changing sentence structure while retaining the meaning, testing the model's ability to maintain accurate predictions despite different grammatical forms. Semantic shifts involved subtle modifications that could alter the entailment relationship between the premise and hypothesis, allowing us to evaluate the model's sensitivity to nuanced meaning changes, such as shifts from entailment to contradiction. **Figure 5** shows that the resulting dataset was more balanced in label distribution compared to the premise-altered contrast set.

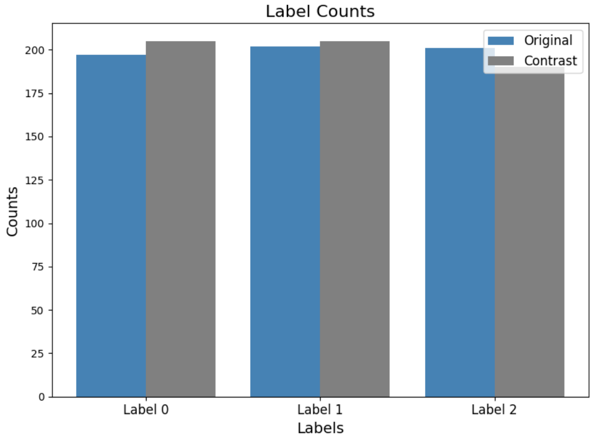


Figure 5: Label distribution, Original data vs. Hypothesis contrast set

3 ELECTRA-small Performance

3.1 SNLI Validation Set Results

The ELECTRA-small model, initialized from the Hugging Face model hub, was pretrained and subsequently fine-tuned on 550,152 examples from the SNLI training set. Following this, the model was evaluated on 10,000 validation examples from the SNLI dataset, with initial results shown in Table 1. The model achieved an accuracy of 89.3%, effectively classifying examples as entailment, contradiction, or neutral.

Evaluation Loss	Evaluation Accuracy
0.3874	0.8929

Table 1: Baseline results

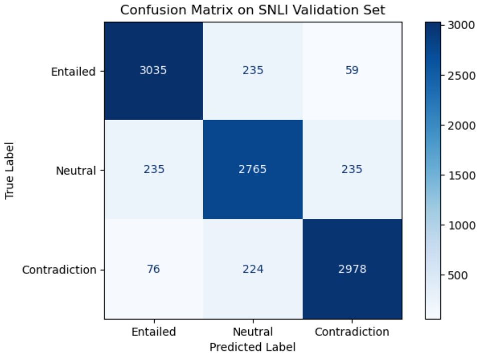


Figure 6: Confusion matrix of trained ELECTRA-small on SNLI validation set

Figure 6. presents a confusion matrix illustrating the classification performance. The diagonal cells represent correctly classified examples, while off-diagonal cells indicate misclassifications. Notably, the model rarely confused contradictory hypotheses with entailment (76 instances) or vice versa (59 instances). However, it struggled more with neutral labels—assigning 235 neutral labels incorrectly to entailment or contradiction examples, and assigning neutral labels incorrectly to 235 entailment and 224 contradiction instances. This suggests that the model's certainty is lower when dealing with neutral relationships, highlighting an area for further improvement.

3.2 Contrast Set Evaluation

The ELECTRA-small model was evaluated on a contrast set consisting of 600 premises altered from the SNLI validation dataset. As expected, performance on the contrast set was lower than on the original dataset, reflecting the model's reliance on detecting dataset-specific artifacts rather than true language understanding. The model achieved an accuracy of 65%, indicating difficulty in handling out-of-distribution (OOD) examples (Swayamdipta et al., 2020).

Evaluation Loss	Evaluation Accuracy
0.7025	0.8233

Table 2: Perturbed Hypothesis Contrast Set results

Subsequently, a contrast set involving 600 altered hypotheses was used, resulting in an accuracy of 82.3% (as shown in Table 2). This higher accuracy on altered hypotheses suggests that the model is more robust to changes in the hypothesis, aligning well with typical evaluation scenarios for natural language inference tasks. Of the 106 samples predicted wrong, 37 of them were semantic shifts, 39 were syntactic rephrasing and 39 were synonym substitution changes The significant accuracy gap between altered premises (65%) and altered hypotheses (82.3%) implies that the model struggles more with changes to the premise, likely due to reliance on dataset-specific patterns. The confusion matrix for the contrast set, presented in **Figure 7**, shows similar trends to the SNLI validation set, with most incorrect predictions involving the neutral class. The off-diagonal values—all of which represent incorrect predictions—are much smaller than the values on the diagonal. Of the off-diagonal values, values at the north, south, and west positions are larger than those in the southwest and northeast positions. Interestingly, there were fewer mistaken contradiction predictions for truly neutral examples in the contrast set compared to the original validation set, indicating some improvement in handling nuanced entailment relationships.

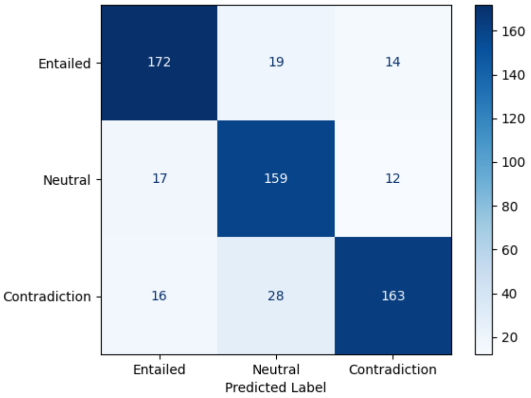
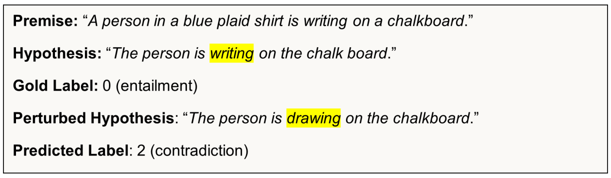


Figure 7: Confusion matrix of trained ELECTRA-small on contrast set. Y-axis: True Label, X-axis: Predicted Label

Figure 8 is an incorrectly labeled example from the hypothesis perturbed contrast subset where synonym substitution was implemented. The word ”Writing” was replaced with ”Drawing,” leading to a prediction of contradiction instead of neutral, despite the gold label being entailment. **Figure 9** presents another incorrect prediction where syntactic rephrasing altered the sentence structure, resulting in a prediction of neutral when the correct label was contradiction. In **Figure 10**, a semantic shift led to an incorrect classification where adding a single word changed the meaning of the hypothesis, and the model incorrectly predicted that the premise supported the hypothesis, though it should have been labeled as neutral.



Premise: “*A person in a blue plaid shirt is writing on a chalkboard.*”

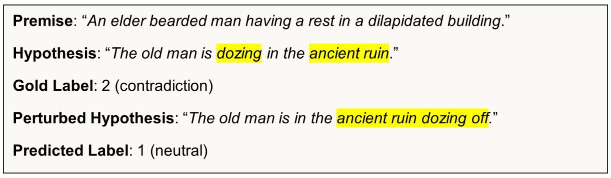
Hypothesis: “*The person is writing on the chalk board.*”

Gold Label: 0 (entailment)

Perturbed Hypothesis: “*The person is drawing on the chalkboard.*”

Predicted Label: 2 (contradiction)

Figure 8: Perturbed Hypothesis: Synonym Substitution



Premise: “*An elder bearded man having a rest in a dilapidated building.*”

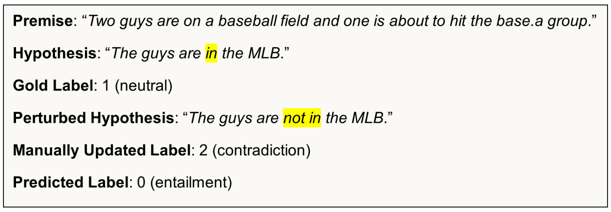
Hypothesis: “*The old man is dozing in the ancient ruin.*”

Gold Label: 2 (contradiction)

Perturbed Hypothesis: “*The old man is in the ancient ruin dozing off.*”

Predicted Label: 1 (neutral)

Figure 9: Perturbed Hypothesis: Syntactic Rephrasing



Premise: “*Two guys are on a baseball field and one is about to hit the base.a group.*”

Hypothesis: “*The guys are in the MLB.*”

Gold Label: 1 (neutral)

Perturbed Hypothesis: “*The guys are not in the MLB.*”

Manually Updated Label: 2 (contradiction)

Predicted Label: 0 (entailment)

Figure 10: Perturbed Hypothesis: Semantic Shift

3.3 Dataset Cartography

To enhance the robustness of the ELECTRA-small model on the SNLI dataset, Dataset Cartography was employed to refine the training data. This method enhances dataset quality by mapping examples based on model behavior during training, categorizing data into easy-to-learn, hard-to-learn, and ambiguous examples by monitoring model confidence and variability throughout training.

Coordinates for data maps were constructed using training dynamics, specifically leveraging the mean and standard deviation of the gold label probabilities predicted for each example across training epochs. These metrics, referred to as confidence and variability, allow for the categorization of examples (Swayamdipta et al., 2020).

The data map shown in **Figure 11** reveals three distinct regions in the dataset:

- **Easy-to-Learn**: Instances with high confidence and low variability, which the model predicts correctly and consistently (red region in Figure 11, top-left).
- **Hard-to-Learn**: Instances with low confidence and low variability, which are rarely predicted correctly (blue region in Figure 11, bottom-left).
- **Ambiguous**: Examples with high variability, indicating model indecision (green region in Figure 11, right-hand side). These examples are crucial for evaluating the model’s decision boundary and generalization capabilities.

The majority of instances belong to the high confidence and low variability region, categorized as easy-to-learn, as they are consistently predicted correctly with high confidence. A smaller group is characterized by low confidence and low variability, often misclassified and referred to as hard-to-learn. The third group contains ambiguous examples, which have high variability, indicating the model’s indecision about their labels.

To select data points for retraining, Dataset Cartography mapped and categorized a 25,000-example subdataset from the original 550,152-example training SNLI dataset. The ambiguous zone was defined as having variability above 0.2, regardless of confidence level. Hard-to-learn examples had confidence below 0.3 and variability less than 0.2. Easy-to-learn examples were defined as having confidence above 0.8 and variability less than 0.2. This process resulted in 19,928 easy-to-learn datapoints, 758 hard-to-learn datapoints, and 2,828 ambiguous datapoints.

4 Fine-Tuning

Metric	Evaluation Loss	Evaluation Accuracy
SNLI Baseline	0.3851	0.8931
Contrast Set	0.6981	0.8233

Table 3: Fine-tuned ELECTRA-small results on selected SNLI validation data

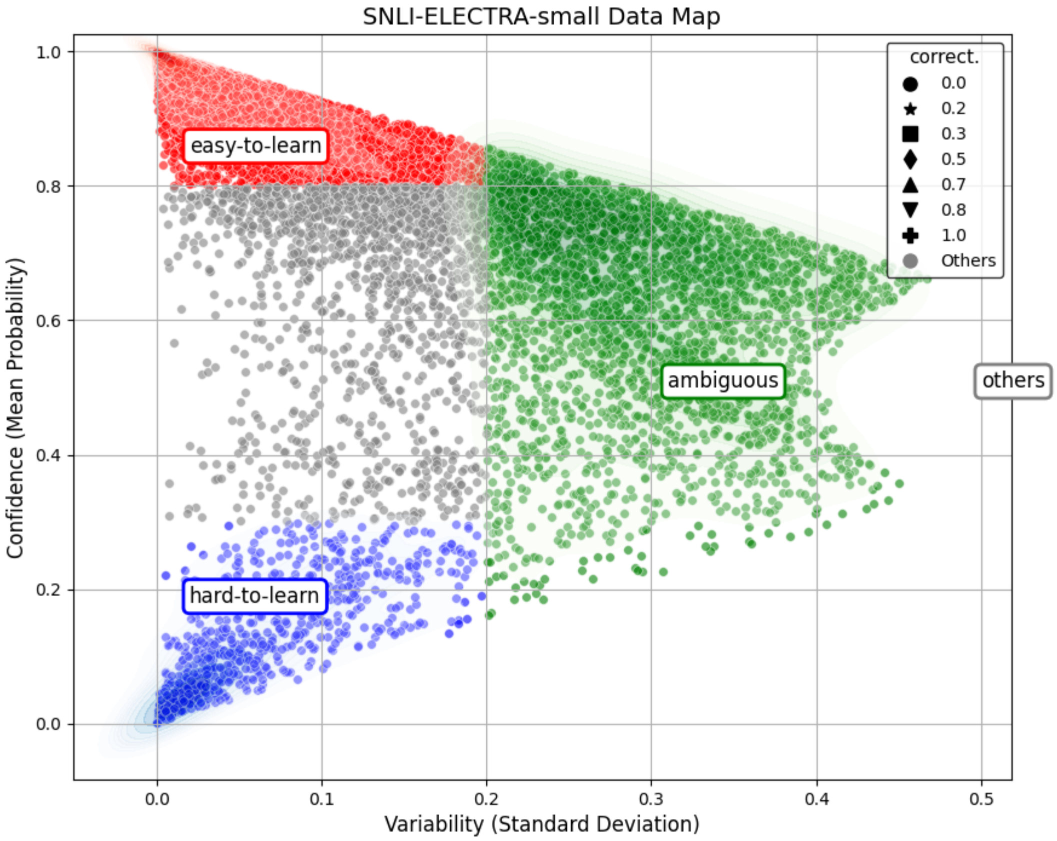


Figure 11: Dataset cartography map for SNLI based on ELECTRA-small. The map follows a bell-shaped curve With respect to model confidence and variability.

4.1 Retraining and Evaluation

A total of 3,000 examples were sampled from the SNLI training set, 20% of which are classified as hard-to-learn, 20% as easy-to-learn, and 60% as ambiguous. The ELECTRA-small model was further fine-tuned on these 3,000 data points, by reducing the learning rate from 5e-05 to 1e-6 and the number of epochs was reduced from 3 to 1. This model was then evaluated against the full SNLI validation set.

Table 3 shows the results from this validation run. The retrained model accurately classified 89.31% of the original SNLI validation dataset. Comparing the results to **Table 1**, there is a slight decrease in loss (0.3851 vs. 0.3874) and very small increase in accuracy (89.31% vs. 89.29%). The retrained model performed similarly to our hypothesis contrast set in loss (0.6981 vs. 0.7025) and in accuracy (82.33% vs 82.33%). Our strategy did not vastly improve the performance of the ELECTRA-small model on the baseline SNLI validation set nor our perturbed hypothesis contrast set.

4.2 Conclusions

We presented a detailed evaluation of the ELECTRA-small model on the SNLI dataset, focusing on its ability to handle nuanced linguistic variations through contrast sets and dataset cartography. Our findings revealed that the model, while effective on the original dataset, struggled with subtle perturbations, particularly in modified premises. The use of dataset cartography enabled the identification of easy-to-learn, hard-to-learn, and ambiguous examples, allowing us to target challenging subsets during retraining. Despite these efforts, the retraining process yielded only marginal improvements in overall model performance, suggesting that fine-tuning on specific subsets does not necessarily overcome the model's reliance on superficial heuristics.

Further research could focus on refining dataset cartography techniques to better differentiate between truly ambiguous examples and instances influenced by annotation artifacts. Exploring alternative retraining strategies, such as curriculum learning or adversarial training, may also enhance model robustness and generalization. By addressing these limitations, the findings from this study can provide valuable insights into improving the reliability and interpretability of transformer-based models in natural language inference tasks.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Empirical Methods in Natural Language Processing (EMNLP)*, 2015:632–642.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *International Conference on Learning Representations (ICLR)*.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Manish Sanwal. 2024. Evaluating large language models using contrast sets: An experimental approach. *International Journal of Artificial Intelligence Research and Development (IJAIRD)*, 2(2):90–97.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *Empirical Methods in Natural Language Processing (EMNLP)*, 2020:9275–9293.