Enhancing NLI Robustness: Leveraging Dataset Cartography and Distractors in Training

James Fu
The University of Texas at Austin
jamesfu@utexas.edu

Abstract

This project explores the application of LLMs to create novel contrast sets with syntactic distractors and reproduces recent research efforts in dataset cartography to systematically characterize and improve model robustness in NLI tasks. By strategically classifying data points into easy-to-learn, hard-to-learn, and ambiguous categories, we can fine-tune their proportions in the training data to mitigate model sensitivity to dataset artifacts via generalization. Whereas the baseline model achieved an F1 score of 89.2% on the unmodified SNLI dataset, the fine-tuned ELECTRA-small model introduced a slight 0.1% improvement, but on a dataset containing hard-to-learn examples and distractors. More notably, our training approach led to an improved performance on a novel contrast set from 56.9% to 65.7%. These results indicate that, in the presence of distractors, rebalancing the proportions accordingly in the dataset training methodology can result in more robust Natural Language Inference (NLI) models that are less sensitive to dataset artifacts and outperform baseline measures.

1 Introduction

Natural Language Inference (NLI) is a core task in Natural Language Processing (NLP) that involves categorizing the semantic relationship between a premise and a hypothesis as entailment, contradiction, or neutrality. Despite the accuracy of various pre-trained models, recent studies have shown that these models often exploit dataset artifacts and lack a true understanding and ability to perform task-specific reasoning, leading to limited generalization capabilities (Poliak et al., 2018).

To test these limitations, a common method is the introduction of contrast sets (Gardner et al., 2020). In short, these sets are modified versions of existing examples in a dataset, designed to challenge models by introducing variations that reveal examples that cause a model to fail. Examples of perturbations include altering word order, paraphrasing, or introducing new semantic constructs. Another common approach is adversarial training, which involves intentionally crafting examples to "fool" the model by altering inputs and labels (Ivgi and Berant, 2021). Although this may seem suitable in this case, our approach instead focuses on introducing subtle perturbations to the hypothesis while keeping the label intact. Specifically, we use LLMs to create a synthetic contrast set containing paraphrased hypotheses and test it on our base model. Implementing LLMs enables more advanced paraphrasing compared to methods that simply introduce syntactic disruptors or retain the same structure while only changing one or two words, which fails to challenge the robustness of our model.

We then reproduce recent data cartography efforts to analyze training dynamics and classify data points into groups such as easy-to-learn, hard-to-learn, and ambiguous using the mean and standard deviation of the gold label probabilities. These classifications enable us to reweigh groups in the training data and fine-tune our initial model on the more challenging examples. This process, known as inoculation by fine-tuning (Liu et al., 2019), allows us to investigate whether performance gaps arise from inherent limitations in the original training data or weaknesses in the model itself. We then evaluate the ELECTRA-small model to determine whether its performance improves after fine-tuning with the more difficult examples and the introduced contrast set.

2 Methodology

2.1 Datasets

To perform NLI, we used the Stanford Natural Inference (SNLI) Corpus (Bowman et al., 2015), which consists of around 570,000 annotated sentence pairs categorized as entailment, contradiction,

or neutral. The dataset includes 550,000 examples for training and 10,000 examples each for validation and testing.

Next, a contrast set containing 10,000 randomly chosen and modified "distractor" hypotheses from the SNLI corpus was generated using LLM prompting. Specifically, we used OpenAI's API and prompted GPT-4o-mini to modify only the hypothesis through paraphrasing, while the original premises and labels remained unchanged (see Figure 1).

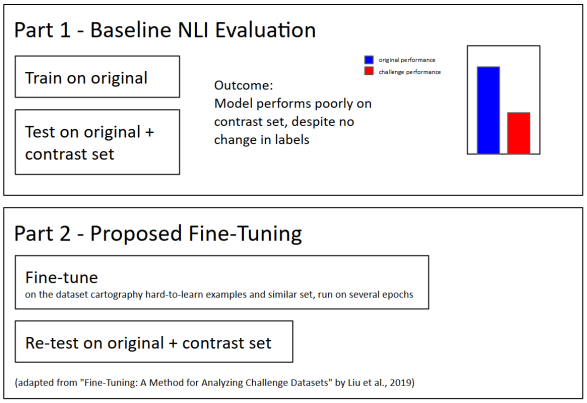

Figure 1: Confidence histogram for challenge dataset examples (adapted from (Liu et al., 2019)).

Table 1 below provides several examples of how our GPT-4o-mini code modifies the hypotheses. While the original label remains unchanged, the base model described in Section 2.2 may predict a different label following these modifications. The specific prompting and code used for generating this contrast set is available online.[1]

Premise	Original Hypothesis	Hypothesis with Distractor
A man is giving a presentation in front of a crowd.	The man is at a sales conference.	The man is at a business seminar.
A man walking down a pathway ending at a lake.	A man is walking outdoors.	A man is walking beside a park.
A boy and a girl are standing near the water on a beach.	A boy and a girl are standing beside the water.	A family enjoys a picnic on the beach, with waves gently lapping at the shore nearby.

Table 1: Examples of original and distractor hypotheses generated by GPT-4o-mini.

[1] github.com/[redacted]/distractor_gen.py

2.2 Model

Here, we begin with a pre-trained ELECTRA-small base model (Clark et al., 2020) fine-tuned on NLI for 3 epochs. This model consists of 12 layers, a hidden size of 256, an FFN inner layer hidden size of 1024, 4 attention heads of size 64, an embedding size of 128, and 14M parameters. After training for 3 epochs, the internal weights of the model are stored, making it reusable without requiring retraining.

As shown in Table 2, the model is able to achieve high accuracy across the three categories on the SNLI test set. The diagonal entries represent true positives, which constitute the majority of classifications.

	Entail	Neutral	Contradict
Entail	3022	240	67
Neutral	214	2781	240
Contradict	73	231	2974

Table 2: SNLI Baseline Confusion Matrix. Columns correspond to predicted values, and rows correspond to actual values.

However, when evaluated with a contrast set, the model struggles with both the "contradict" and "neutral" categories, resulting in high misclassification rates of 54.98% and 54.10%, respectively. The model exhibits high rates of confusion between "contradict" and "entail", as well as between "neutral" and "entail", leading to a high rate of error across these categories.

	Entail	Neutral	Contradict
Entail	2641	431	219
Neutral	1567	1538	247
Contradict	1444	402	1511

Table 3: SNLI Contrast Set Confusion Matrix. Columns correspond to predicted values, and rows correspond to actual values.

Overall, when perturbed with a contrast set the model displays a clear tendency to conflate "entail" with both "contradict" and "neutral". Further analysis of general error categories will be brought up in the discussion section.

2.3 Dataset Cartography

For the second part of our experiment, we utilized a training dynamics approach following the re-

search methodology proposed by (Swayamdipta et al., 2020a). Confidence measures the average probability assigned to the correct label, variability tracks fluctuations in the probability label in different epochs, and correctness is a discrete model reflecting how accurate the model is. In our case, we track these metrics across three epochs of model training, enabling us to capture any variation in predictions.

The scatter plot in Figure 2[2] visualizes our data map for the SNLI dataset using the ELECTRA-small model, plotting confidence against variability. Using the Plotly library, I referenced the AllenAI implementation of data cartography (Swayamdipta et al., 2020b)[3] to create a heatmap-style visualization, where data points are color-coded by confidence, with darker colors representing higher confidence.

After implementation, data cartography categorizes data points as easy-to-learn, ambiguous, or hard-to-learn, providing insights into dataset quality and its impact on model performance. The HuggingFace tokenizer allows us to identify instance IDs corresponding to hard-to-learn data, enabling adjustments to the proportion of such data included in our training sets.

3 Discussion

3.1 Fine-Grained Error Categories

Earlier, we discussed how the base ELECTRA-small model struggles to distinguish between the neutral and contradiction classes. In this context, examining the overlap between the contrast set and the SNLI evaluation set is likely the most effective way to understand why the model performs poorly on the contrast set. Although the premises in the contrast set and the SNLI evaluation set have minimal overlap ($n < 30$ overlapping examples), the first semantic shift correction category uses limited examples to analyze why paraphrasing leads to erroneous classifications. The subsequent categories instead compare the original and modified hypotheses to identify why the model changed its label and made an erroneous prediction.

[2]Due to the size of the data map, it was placed near the end of the document.

[3]Reference code used to create data maps: https://github.com/allenai/cartography.

3.1.1 Semantic Shift Correction of Overlapped Examples Leading Neutral Examples to Entailment or Contradiction

These errors arise during paraphrasing when the LLM unintentionally alters semantic relationships to create the contrast set. This leads the contrast set to inadvertently mark the example with the wrong label, despite the fact that the model correctly identifies the relationship given the revised context and semantic meaning.

For instance, consider the premise *"A man in a wetsuit surfing"* and the original hypothesis *"The surfboard the man is on is yellow,"* where the latter of the two is paraphrased as *"A person is riding a wave"* when creating the contrast set. The golden label in the unmodified SNLI test set implies neutrality, which is correct since the color of the surfboard cannot be derived from the premise. However, after paraphrasing the contrast set hypothesis to *"A person is riding a wave",* the semantic relationship and specifics about color needed to classify the golden label are lost. In short, these changes cause the original and contrast labels to differ, even though the model correctly identifies the relationship in the revised context. This leads errors to arise when our paraphrase method unintentionally shifts meaning and alters the label and causes the performance of the model to drop significantly when tested under the contrast set.

3.1.2 Synonym Replacement or Verb Substitution Induced Ambiguity

These errors occur when paraphrasing replaces specific terms with more ambiguous synonyms, reducing the semantic precision of the original hypothesis. While the implementation of contrast sets was intended for the premise to maintain equivalence with the original hypothesis, the paraphrasing introduces vagueness, causing the model to interpret the example as neutral.

For example, the premise *"A musician is playing an instrument on a stage"* and the original hypothesis *"The professional pianist is from Asia and is ready to perform,"* are initially correctly labeled as entailment. However, the contrast set paraphrases the hypothesis to *"A musician is on stage getting ready."* The shift in meaning between *"getting ready"* and *"playing an instrument"* introduces a clear difference, changing the label to neutral and leading the model to misclassify the example. The paraphrased hypothesis fundamentally alters the

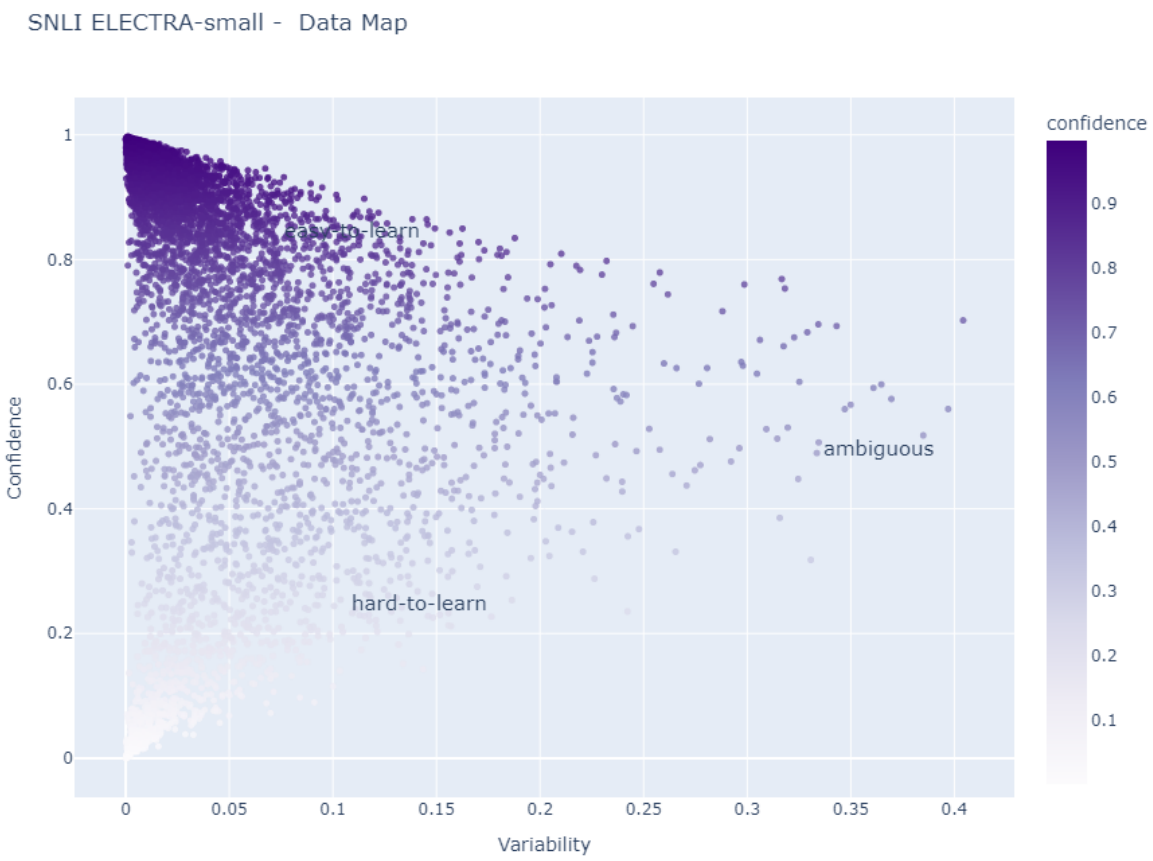

Figure 2: Heatmap of SNLI Dataset Examples Using ELECTRA-Small Model

relationship, focusing on preparation rather than performance.

3.1.3 Inference Scope Expansion Leading to Neutrality during Space-Time References

These errors often happen when dealing with time or places, and paraphrasing introduces inferences or conclusions that make additional assumptions beyond the information provided in the original hypothesis. The implementation of contrast sets was intended to preserve equivalence, yet the paraphrased hypotheses often add unsupported details, causing the model to classify the relationship as neutral.

One example consists of the premise *"A man walking down a pathway ending at a lake,"* and the original hypothesis *"A man is walking beside a park,"* which is correctly labeled as entailment since a pathway ending at a lake is likely located within or near a park. However, the contrast set paraphrases the hypothesis to *"A man is walking to a cabin."* The introduction of *"a cabin"* does not necessarily imply that the man is *"beside a park,"* leading to unintended consequences due to an unsupported assumption. The paraphrased hypothesis fundamentally changes the original hypothesis and introduces ambiguity.

3.2 Fine-Tuning Results with Data Cartography

As outlined earlier, our goal was to train the model on the original training set and then fine-tune on hard-to-learn examples for many (>3) epochs. However, due to limitations in computing power and constraints in our coding environment, we were unable to fully implement the inoculation method proposed by (Liu et al., 2019). Instead, we trained the model for three epochs and subsequently retrained it using a subset of challenging examples that the model initially struggled with. These hard-to-learn examples were selected based on the data map generated in Section 2.3, employing defined thresholds: confidence levels below 0.5, correctness below 0.6, and variability exceeding 0.35.

Compared to the initial model trained for three

epochs on the SNLI test set, this refined model's training dataset places an increased weight on harder-to-learn examples, which were comparatively included twice as often. Additionally, newly generated contrast sets, along with a similar but separate contrast set containing distractors, were introduced to further "boost" the NLI abilities of the model.

Table 4 below shows the evaluation metrics before and after these modifications. Although this approach only resulted in a negligible 0.1% improvement on the best-performing dataset (marked with an asterisk) of the fine-tuned model, it significantly enhanced performance on a contrast set. This demonstrates that increasing the weight of hard-to-learn samples through data cartography is a reproducible and effective strategy.

	ELECTRA-small		Fine-tuned ELECTRA	
Dataset	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Baseline	*89.2%	89.2%	88.9%	88.6%
Contrast	56.9%	56.5%	65.7%	65.7%
Full	51.2%	51.3%	*89.3 %	89.4%

Table 4: Accuracy and F1 scores across different training methods. The first and second columns show the performance of the ELECTRA-small model on the baseline and fine-tuned models, respectively. The three datasets evaluated are: (i) the unmodified SNLI training set (baseline), (ii) a contrast set (contrast), and (iii) a combined dataset consisting of the contrast set, hard-to-learn examples, and the unmodified SNLI training set (full).

Alongside accuracy metrics, F1 scores are included to provide a more in-depth analysis, since neutral and contradiction examples may be overrepresented in the reweighed dataset. Recall that our confusion matrix found that the model was more likely to mislabel these examples, causing them to appear more frequently than their entailment counterparts in the modified training set derived from dataset cartography.

4 Conclusion

A future improvement for addressing errors could involve fine-tuning the prompting or providing annotators with guidelines to ensure that paraphrases do not alter the intended meaning of the original hypothesis. However, this may be challenging to achieve without using predisposed knowledge of the golden label as a basis for comparison. Our analysis of error categories agrees with our confusion matrix, since both exemplify labels being shifted into the predicted neutral bin, despite not being truly neutral.

Our project exemplifies that the implementation of contrast sets and data cartography can improve the accuracy by training with a higher proportion of challenging examples. This approach also enhances the baseline accuracy of the full model by 0.1%, with a minimal 0.3% reduction in the evaluation metrics of the original SNLI test set when assessed using the newly fine-tuned model. Most importantly, the accuracy when evaluated on a novel contrast set increased by roughly 8%. Further improvements can be made while repeating the same training and fine-tuning process with the ELECTRA-small model by refining the contrast set to target and address semantic shift errors.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Maor Ivgi and Jonathan Berant. 2021. Achieving model robustness through discrete adversarial training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1529–1544, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020a. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Swabha Swayamdipta et al. 2020b. Data cartography code. https://github.com/allenai/cartography.