Investigating Data Efficient Methods for Enhancing Model Generalization using Dataset Cartography, Augmentation and Perturbation Techniques

Kevin Rohling

kevin@kevinrohling.com
https://kevinrohling.com
https://github.com/krohling
https://www.linkedin.com/in/krohling/
The University of Texas at Austin, Masters in AI, Department of CS

Abstract

This work investigates data-efficient strategies for improving generalization using Dataset Cartography [Swayamdipta et al., 2020], augmentation, and perturbation-based uncertainty quantification techniques. We begin by establishing baseline performance using the `google/electra-small-discriminator` [Clark et al., 2020] model trained on SQuAD [Rajpurkar et al., 2016]. We then extend the traditional Dataset Cartography calculation of intra-sample confidence and variability, derived from training dynamics, to also include inter-sample confidence/variability using Sampling with Perturbation for Uncertainty Quantification (SPUQ) [Gao et al., 2024], allowing us to also capture a measure of epistemic uncertainty. Both metrics are then used to map training examples in terms of their variability, allowing us to select subsets of examples for training, focusing on those with high variability. These subsets are tested with multiple data augmentation strategies and used to train separate models, which are evaluated on both in-domain and out-of-domain performance. In total we evaluate 20 different experimental configurations, including training on subsets selected by intra-sample, inter-sample, and aggregated variability, with and without data augmentation.

We find that while training on the full dataset yields the strongest in-domain performance, targeting subsets of highly variable (ambiguous) examples outperforms random subsets of equivalent size on both in-domain and out-of-domain evaluations, with the most significant gains observed in the latter. Data augmentation alone did not improve in-domain performance but did result in performance improvements for out-of-domain evaluation. The most substantial improvements were achieved by training on subsets selected by aggregated variability, which combines intra-sample and inter-sample uncertainty metrics, and augmenting with both paraphrased and adversarial examples. These findings suggest that focusing on challenging examples and augmenting training data is a data efficient method for enhancing model robustness and generalization to out-of-domain distribution shifts.

1 Methodology

Dataset Cartography Swayamdipta et al. [2020] introduced *Dataset Cartography*, a framework for analyzing training dynamics to identify and categorize dataset examples based on the model's behavior during training. Specifically, they track two metrics for each training example over multiple epochs:

- **Confidence** (μ_i): The mean probability assigned to the true label by the model across training epochs.
- **Variability** (σ_i): The standard deviation of the model's predicted probabilities for the true label across epochs.

By plotting examples based on their confidence and variability, Dataset Cartography identifies three regions:

1. **Easy-to-Learn**: High confidence and low variability, examples the model consistently predicts correctly.
2. **Ambiguous**: Moderate confidence and high variability, examples where the model's predictions fluctuate during training.
3. **Hard-to-Learn**: Low confidence and low variability, often associated with mislabeled or difficult to learn examples.

For each training example i, we tracked the model's predicted probability for the correct answer over multiple epochs, calculating:

$$C_{\text{intra}}(i) = \mu_i = \frac{1}{E} \sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^* | x_i) \tag{1}$$

$$Var_{\text{intra}}(i) = \sigma_i = \sqrt{\frac{1}{E} \sum_{e=1}^{E} \left(p_{\theta^{(e)}}(y_i^* | x_i) - \mu_i\right)^2} \tag{2}$$

where $p_{\theta^{(e)}}(y_i^*|x_i)$ is the model's predicted probability for the correct answer at epoch e, μ_i is the mean intra-sample confidence, and σ_i is the standard deviation (intra-sample variability).

Sampling with Perturbation for Uncertainty Quantification (SPUQ) While Dataset Cartography uses training dynamics to assess model confidence and variability, it primarily captures *aleatoric uncertainty*, which arises from inherent data ambiguity. To incorporate *epistemic uncertainty*—uncertainty due to model limitations—Gao et al. [2024] proposed *Sampling with Perturbation for Uncertainty Quantification* (SPUQ). SPUQ extends traditional uncertainty quantification by introducing input perturbations and sampling methods.

The inter-sample confidence for a target example is calculated as:

$$C_{\text{inter}}(i) = \frac{\sum_{j \in P(i)} w_{ij} \cdot s(y_i, y_j)}{\sum_{j \in P(i)} w_{ij}} \tag{3}$$

where $P(i)$ is the set of perturbed examples for example i, $s(y_i, y_j)$ is a similarity score between the predicted answers for i and j, and w_{ij} is a weight based on the similarity between the contexts of i and j. Following the recommendations of Gao et al. [2024], we use Rouge-L as a similarity score, $s(y_i, y_j)$, and 5 input perturbations were generated (via paraphrasing) for each dataset example.

Aggregated Uncertainty In addition to evaluating intra-sample and inter-sample uncertainties separately, we calculate an aggregated confidence metric as the mean of both. This allowed us to evaluate training on subsets of examples with both high intra-sample and high inter-sample uncertainties. To combine intra-sample and inter-sample uncertainties, we calculate the following aggregated confidence metric:

$$C_{\text{agg}}(i) = \frac{C_{\text{intra}}(i) + C_{\text{inter}}(i)}{2} \tag{4}$$

2 Datasets

SQuAD Dataset (In-Domain Training/Evaluation) The Stanford Question Answering Dataset (SQuAD) [Rajpurkar et al., 2016] is a widely used benchmark for evaluating machine reading comprehension models. It consists of question-answer pairs derived from Wikipedia articles where each question is associated with a context paragraph, and the answer is a span of text within the context.

Adversarial QA Dataset (Out-Of-Domain Evaluation) The AdversarialQA dataset [Bartolo et al., 2020] is designed to evaluate the robustness of QA models by providing adversarial examples. This dataset contains examples with intentional distractor content that are designed to be more difficult for models to answer correctly. AdversarialQA is used in this project to assess the model's generalization capabilities on out-of-domain data.

Paraphrased Augmentation Dataset (`squad_paraphrasing_5`) To introduce additional diversity in the phrasing of SQuAD training examples, as well as aid in the calculation of SPUQ scores, we generated a paraphrased version of the SQuAD dataset, referred to as `squad_paraphrasing_5`. For each original question and context pair, we created five paraphrased augmentations resulting in a total of 522k paraphrased examples.

Adversarial Augmentation Dataset (`squad_adversarial`) To improve model robustness, we generated a dataset of adversarial augmentations, named `squad_adversarial`. For each entry in the SQuAD training dataset, we created an adversarial example by introducing distracting or misleading information into the context. This dataset contains 174k adversarial examples.

Data Augmentation Strategy To generate the augmentation datasets, we used OpenAI's GPT-4o-mini model to produce both the paraphrased and adversarial datasets. The model was prompted via the OpenAI API with custom templates for each augmentation type. These prompts were designed to generate paraphrased and adversarial examples while preserving the original answer spans. Every example in the training portion of the original SQuAD dataset was added to the prompt and sent to GPT-4o-mini. The resulting paraphrased and adversarial examples were added to the respective augmentation datasets resulting in a total of 696k augmentation examples.

3 Experiment Configurations

We conducted a series of experiments to evaluate the impact of both data augmentation and focused training on high variability examples. The configuration of these experiments varied along two axes: the subset of the dataset used for training and the application of data augmentation techniques. We used five different subset selections from the SQuAD dataset for training, each representing a different selection strategy. These configurations were designed to establish baselines and evaluate subsets selected using variability metrics. Additionally, we applied four different data augmentation strategies to each subset resulting in a total of 20 unique experimental configurations. Each of these configurations was then used to train a new model which was evaluated on both the in-domain SQuAD dataset and the out-of-domain AdversarialQA dataset after 5 epochs.

Baseline Model Configurations The following dataset configurations were used to train models to use as baselines for comparison. The **Full Dataset** configuration was used to compare models trained on subsets to the performance of a model trained on the full dataset. While equivalent performance was not expected, this configuration served as a hypothesized upper bound for the performance of models trained on smaller subsets. The **Random 33%** subset functioned as a peer baseline for the other subset configurations, providing a more direct comparison point trained on the same number of examples.

Subset Selection Using Variability Metrics The subsets for each configuration were selected by sorting the training examples in descending order based on the specified variability metric (intra-sample, inter-sample, or aggregated) and selecting the top 33%. This approach targets the most ambiguous examples relative to the specified metric. These included the following variations:

- *Top 33% Sorted by High Intra-Sample Variability*
- *Top 33% Sorted by High Inter-Sample Variability*
- *Top 33% Sorted by High Aggregated Variability*

Data Augmentation Strategies Data augmentation was applied using the following configurations:

- *No Augmentation*: No augmentation examples were added to the training set.
- *Paraphrasing Augmentation*: Paraphrased examples were added to the training set.
- *Adversarial Augmentation*: Adversarial examples were added to the training set.
- *Both*: Both paraphrased and adversarial examples were added to the training set.

Evaluation Metrics The models were evaluated using the HuggingFace SQuAD evaluation script, computing exact match (EM) and F1 scores on the validation set of both the SQuAD dataset and the AdversarialQA dataset to assess in-domain and out-of-domain performance, respectively.

4 Results

The experimental results, summarized in Tables 1a–2b, reveal clear patterns with respect to training set composition and data augmentation strategies. We evaluate both in-domain performance on SQuAD and out-of-domain generalization on AdversarialQA. As expected, training on the full dataset achieves the highest in-domain scores. However, restricting training to carefully selected subsets identified as "ambiguous" (i.e., high variability examples) leads to performance improvements when measured against the randomly selected baselines. Additionally, augmenting training data with paraphrased and adversarial examples enhances out-of-domain generalization, particularly when combined with subsets selected for high variability. The most substantial out-of-domain improvements are observed when training on subsets selected by aggregated variability and augmented with both paraphrased and adversarial examples.

Table 1: In-Domain Scores

(a) F1 Scores

Configuration	None	Para.	Adv.	Both
Full Dataset	85.2	85.2	85.2	85.1
Random 33%	80.3	80.6	81.4	81.8
Intra-sample Var. (High)	82.0	81.5	81.6	82.0
Inter-sample Var. (High)	81.1	81.2	80.7	81.9
Aggregated Var. (High)	82.0	81.6	81.1	82.3

(b) Exact Match (EM) Scores

Configuration	None	Para.	Adv.	Both
Full Dataset	77.4	77.2	77.2	77.3
Random 33%	71.2	72.1	72.4	72.5
Intra-sample Var. (High)	73.3	72.5	72.9	73.3
Inter-sample Var. (High)	71.7	71.3	71.2	72.8
Aggregated Var. (High)	72.6	72.6	71.9	73.0

In-Domain Performance Models trained on the full SQuAD dataset without augmentation achieve an F1 score of 85.2 and an EM score of 77.4. Reducing the training data to a random 33% subset unsurprisingly lowers the baseline in-domain F1 to 80.3 and EM to 71.2. By contrast, models trained on subsets chosen according to high intra-sample variability or aggregated variability achieve slightly stronger in-domain performance (e.g., 82.0 F1 and 73.3 EM for intra-sample variability, no augmentation; 82.0 F1 and 72.6 EM for aggregated variability, no augmentation). This indicates that focusing on more "ambiguous" examples—those that challenge the model during training—can partially compensate for using fewer training instances.

Data augmentation alone does not yield improvements over the unaugmented condition in-domain, and in some cases, augmentation slightly reduces the F1 or EM scores. For instance, paraphrasing or adversarial augmentation when training on the full dataset remains around 85.1–85.2 F1 and 77.2–77.3 EM, showing no meaningful gains. Similarly, subsets selected for intra-, inter-, or aggregated variability show minimal in-domain improvements from augmentation. These findings suggest that adding more data through paraphrasing or adversarial methods does not inherently strengthen in-domain performance beyond what can be achieved by careful example selection.

Out-of-Domain Performance A more pronounced effect is observed on the AdversarialQA evaluation, where out-of-domain F1 and EM scores are generally lower. The full dataset condition starts at 28.1 F1 and 18.0 EM. Augmentation slightly improves these out-of-domain scores (up to 29.4 F1 and 19.0 EM with both paraphrasing and adversarial augmentation). Importantly, the variability-based subsets achieve stronger out-of-domain performance than a random subset of the same size, confirming the hypothesis that focusing on ambiguous examples can enhance robustness. For example, training on the top 33% of examples by intra-sample variability yields an F1 of 24.7 without augmentation and rises to 27.0 with both paraphrasing and adversarial augmentation. Similarly, the top aggregated variability subset shows even more pronounced gains: from 24.0 F1 (no augmentation) to 27.9 F1 (both augmentations), and from 14.9 EM (no augmentation) to a maximum of 18.4 EM with both augmentations. This model even outperforms the model trained on the full dataset without augmentation in EM score (18.4 vs 18.0) and approaches the full dataset with both augmentations in F1 score (27.9 vs. 28.1).

5 Conclusion

This study explored targeted strategies for improving model generalization under data-efficient conditions by leveraging Dataset Cartography, perturbation-based uncertainty quantification, and data augmentation techniques. By using uncertainty metrics to identify "ambiguous" examples—those characterized by high intra-sample, inter-sample, or aggregated variability—we were able to extract smaller, more impactful training subsets. Training on these subsets led to gains in model performance relative to randomly selected baselines, with the most significant improvements

Table 2: Out-of-Domain Scores (AdversarialQA)

(a) F1 Scores

Configuration	**None**	**Para.**	**Adv.**	**Both**
Full Dataset	28.1	29.1	27.7	29.4
Random 33%	21.7	23.9	23.6	25.5
Intra-sample Var. (High)	24.7	25.9	25.0	27.0
Inter-sample Var. (High)	25.6	25.2	24.7	26.5
Aggregated Var. (High)	24.0	26.5	25.6	27.9

(b) Exact Match (EM) Scores

Configuration	**None**	**Para.**	**Adv.**	**Both**
Full Dataset	18.0	19.0	17.6	19.0
Random 33%	12.1	14.3	14.2	15.3
Intra-sample Var. (High)	15.4	15.8	15.5	17.5
Inter-sample Var. (High)	15.8	14.6	14.8	16.7
Aggregated Var. (High)	14.9	16.8	16.0	18.4

observed in out-of-domain evaluations. While data augmentation did not enhance in-domain performance, it did improve out-of-domain generalization, particularly when combined with subsets selected for high variability. The most substantial gains were achieved by training on subsets selected by aggregated variability and augmenting with both paraphrased and adversarial examples.

References

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, December 2020. ISSN 2307-387X. URL http://dx.doi.org/10.1162/tacl_a_00338.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020. URL https://arxiv.org/abs/2003.10555.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. Spuq: Perturbation-based uncertainty quantification for large language models, 2024. URL https://arxiv.org/abs/2403.02509.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics, 2020. URL https://arxiv.org/abs/2009.10795.

A Dataset Cartography

We generated Dataset Cartography plots to visualize the training examples based on their confidence and variability metrics using both intra-sample and inter-sample confidence. Figure 1 and Figure 2 illustrate the distribution of training examples, highlighting the regions corresponding to easy-to-learn, ambiguous, and hard-to-learn examples using both metrics. Our key findings include:

Ambiguous Examples Enhance Robustness Subsets selected for their high variability consistently outperformed random subsets of equivalent size in out-of-domain evaluations, confirming that focusing on inherently challenging instances fosters more robust representations. Notably, these targeted subsets helped bridge the performance gap introduced by training on less data overall.

Augmentation Alone Did Not Improve In-Domain Scores Incorporating paraphrased or adversarially perturbed examples did not significantly lift in-domain metrics, which remained best when training leveraged the full dataset. However, adding these synthetic variants did not diminish performance and provided a crucial boost to out-of-domain generalization.

Combining Uncertainty Metrics and Augmentation Yields Strongest Gains raining on subsets defined by aggregated uncertainty—those reflecting both intra-sample (aleatoric) and inter-sample (epistemic) variability—yielded stronger out-of-domain improvements than either metric alone. When combined with both paraphrased and adversarial augmentations, these subsets produced models that approached or exceeded the out-of-domain performance of models trained on the full dataset without augmentation.

Practical Trade-Offs and Considerations While maximum in-domain accuracy still demands comprehensive training data, our results suggest that selectively focusing training on challenging examples, supplemented by strategically generated augmentations, can yield models that generalize better to novel, adversarial, or otherwise domain-shifted inputs. This approach can be particularly valuable in scenarios where data collection is costly or domain adaptation is critical.

In summary, this work demonstrates that uncertainty-aware, data-efficient training strategies can meaningfully improve model resilience to domain shifts. Future research might investigate more sophisticated perturbation methods, other uncertainty metrics, or adaptive augmentation schedules, aiming to further refine and automate the selection of examples that drive robust, transferable NLP models.

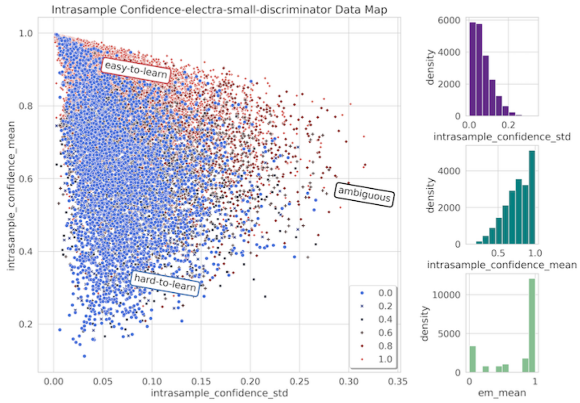

Figure 1: Dataset Cartography Plot of SQuAD using **Intrasample** Confidence and Variability.

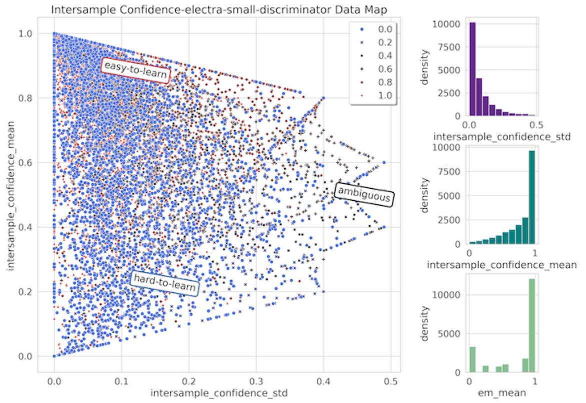

Figure 2: Dataset Cartography Plot of SQuAD using **Intersample** Confidence and Variability.

The Dataset Cartography plots for both metrics demonstrate very similar distributions for dataset examples across the easy-to-learn, ambiguous, and hard-to-learn regions. As can be seen by looking at the *(intra/inter)_sample_confidence_std* and *(intra/inter)_sample_confidence_mean* historgrams, the majority of examples fall into the easy-to-learn category with low variability and high confidence. However, a significant portion of the dataset also falls into the ambiguous region which was the focus for selecting training subsets. Note that this distribution is in agreement with the findings from Swayamdipta et al. [2020].

B Data Augmentation Prompts

B.1 Paraphrasing Augmentation System Prompt

You are being provided with QUESTION, CONTEXT, and ANSWER texts. Your task is to provide {count} paraphrasings of both the QUESTION and CONTEXT. The paraphrasings should maintain the same meaning as the original texts and MUST contain the ANSWER span. If the paraphrased CONTEXT does not contain the ANSWER, it will be invalid. The paraphrased question should be semantically similar to the original but can be rephrased for variety.
Here are a few examples to follow:

Example 1:
QUESTION: ’Who was the first president of the United States?’
CONTEXT: ’George Washington was the first president of the United States.’
ANSWER: ’George Washington’
Paraphrased Question: ’Who became the inaugural president of the United States?’
Paraphrased Context: ’The United States’ first president was George Washington.’

Example 2:
QUESTION: ’When did the Titanic sink?’
CONTEXT: ’The Titanic sank on April 15, 1912, after hitting an iceberg.’
ANSWER: ’April 15, 1912’
Paraphrased Question: ’What year did the Titanic sink?
Paraphrased Context: ’After striking an iceberg, the Titanic sank on the night of April 15, 1912.’

B.2 Adversarial Augmentation System Prompt

You are being provided with QUESTION, CONTEXT, and ANSWER texts. Your task is to provide {count} adversarial versions of the CONTEXT. Adversarial versions should add distractors or misleading information that could confuse a model, while keeping the correct ANSWER in the CONTEXT. Ensure the ANSWER remains accurate in the modified CONTEXT.
Here are a few examples to follow:

Example 1:
QUESTION: ’Where was Albert Einstein born?’
CONTEXT: ’Albert Einstein was born in Ulm, Germany, in 1879.’
ANSWER: ’Ulm, Germany’
Adversarial Context: ’Albert Einstein was born in Ulm, Germany, in 1879. Some believe he was born in Munich, but that is incorrect.’

Example 2:
QUESTION: ’What is the capital of France?’
CONTEXT: ’Paris is the capital of France.’
ANSWER: ’Paris’
Adversarial Context: ’Paris is the capital of France, although some mistakenly think Lyon holds that title.

B.3 User Prompt

QUESTION: {question}

CONTEXT: {context}

ANSWER: {answer}

VERY IMPORTANT: Ensure that the new context contains the ANSWER (’{answer}’). If the answer is not present in the new context, the augmentation will be invalid.