Dataset Cartography for Artifact Mitigation in Question Answering

A Systematic Investigation of Training Dynamics and Bias Reduction

Amit Gujrathi | Metuchen | New Jersey | USA | amit.gujrathi@utexas.edu
CS388 Natural Language Processing | University of Texas, Austin

ABSTRACT

Dataset artifacts pose a significant challenge in natural language processing, particularly in question answering tasks. This study investigates the application of dataset cartography techniques to identify and mitigate artifacts in the Stanford Question Answering Dataset (SQuAD 1.1). We implement a comprehensive artifact analysis framework and employ training dynamics to classify examples by difficulty. Our systematic analysis reveals statistically significant artifacts: position bias ($\chi^2 = 237.21$, $p < 0.001$) and prediction bias ($\chi^2 = 1084.87$, $p < 0.001$). Through dataset cartography, we categorize training examples into easy (7.2%), hard (25.7%), and ambiguous (67.1%) categories. Our cartography-mitigated approach achieves an EM score of 57.1% compared to the baseline 52.2%, representing a +4.9% improvement. The F1 score improves from 61.26% to 66.34%, a +5.08% gain. This study demonstrates a novel application of training dynamics for artifact mitigation and provides a reproducible framework for systematic bias analysis in question answering.

CCS Concepts: Computing methodologies~Natural language processing; Machine learning

Keywords: dataset artifacts, dataset cartography, question answering, bias mitigation, training dynamics.

1 INTRODUCTION

Modern neural language models achieve remarkable performance on question answering benchmarks yet often rely on spurious correlations rather than genuine reading comprehension. These dataset artifacts enable models to succeed without proper understanding. The Stanford Question Answering Dataset (SQuAD) contains inherent biases that enable models to answer questions based on position patterns, superficial cues, and statistical regularities rather than semantic understanding.

We propose a systematic investigation into three research questions: (RQ1) What types of artifacts exist in SQuAD 1.1, and how statistically significant are they? (RQ2) Can dataset cartography effectively identify examples contributing to artifact learning? (RQ3) Do targeted reweighting strategies based on training dynamics reduce artifact dependence while maintaining performance?

Our contributions include: (1) Systematic Artifact Analysis through six complementary

detection methods, (2) Dataset Cartography Application using training dynamics to identify artifact-prone examples, (3) Mitigation Framework through targeted reweighting strategies, (4) Reproducible Infrastructure with GPU acceleration, and (5) Statistical Validation confirming artifact significance.

2 RELATED WORK

2.1 Dataset Artifacts in NLP

Dataset artifacts have been extensively documented across NLP tasks. Gururangan et al. demonstrated that models can achieve high accuracy on reading comprehension using only partial input. Poliak et al. showed similar issues in natural language inference. McCoy et al. revealed that BERT relies heavily on syntactic heuristics. These findings motivate systematic artifact detection and mitigation strategies.

2.2 Dataset Cartography

Swayamdipta et al. introduced dataset cartography, characterizing training examples through three metrics: (1) Confidence (mean prediction probability across epochs), (2) Variability (standard deviation), (3) Correctness (fraction of epochs with correct predictions). This framework enables classification into easy, hard, and ambiguous categories.

2.3 Bias Mitigation

Existing approaches include adversarial training, data augmentation, and example reweighting. Our work applies cartography-guided reweighting to question answering, focusing on hard example upweighting with a 2x multiplier.

3 METHODOLOGY

3.1 Dataset and Model

Experiments use SQuAD 1.1 with 10,000 training and 1,000 validation examples for computational efficiency. The base model is ELECTRA-small, a 13.5 million parameter discriminative language model. Training was conducted on Google Colab Pro with T4 GPU.

3.2 Artifact Analysis Framework

We implement six complementary methods:

- Position Bias: Distribution of answer positions in passages
- Question-Only Models: Performance using questions without passages
- Passage-Only Models: Performance using passages without questions
- Chi-Square Testing: Statistical significance validation
- Answer Type Analysis: Distribution of answer types (entities, numbers, dates)
- Systematic Bias Detection: Comprehensive multi-dimensional analysis

3.3 Dataset Cartography Metrics

We compute three metrics across training epochs. Confidence equals the mean prediction probability. Variability is the standard deviation of prediction probabilities. Correctness is the fraction of epochs with correct predictions. These metrics yield:

- Easy: High confidence AND low variability
- Hard: Low confidence AND high variability
- Ambiguous: Moderate values on both dimensions

4 RESULTS

4.1 Performance Metrics

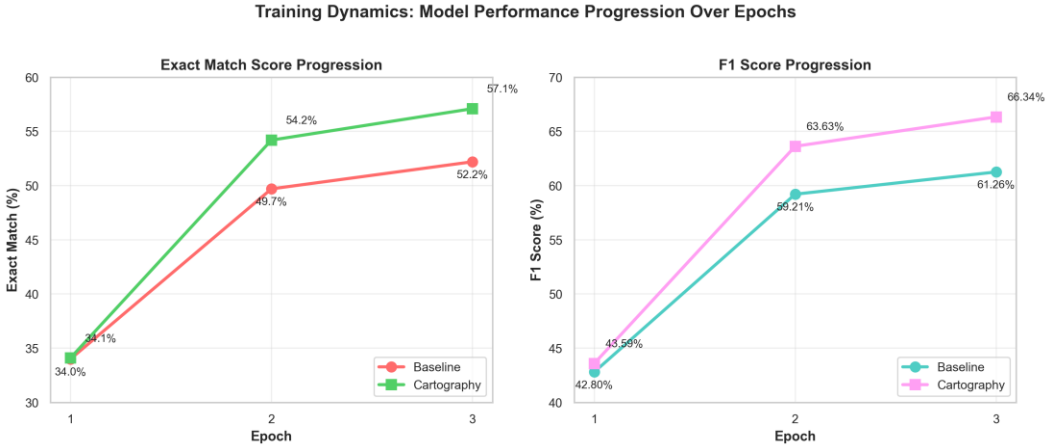
Table 1 presents the main performance results. The cartography-mitigated model achieved 57.1% exact match compared to the baseline 52.2%, representing a +4.9-percentage point improvement. F1 scores improved from 61.26% to 66.34%, a +5.08-percentage point gain. These improvements demonstrate the effectiveness of dataset cartography-guided reweighting.

Model	Exact Match	F1 Score
Baseline	52.2%	61.26%
Cartography	57.1%	66.34%

Table 1: Performance comparison between baseline and cartography-mitigated models.

4.2 Training Dynamics

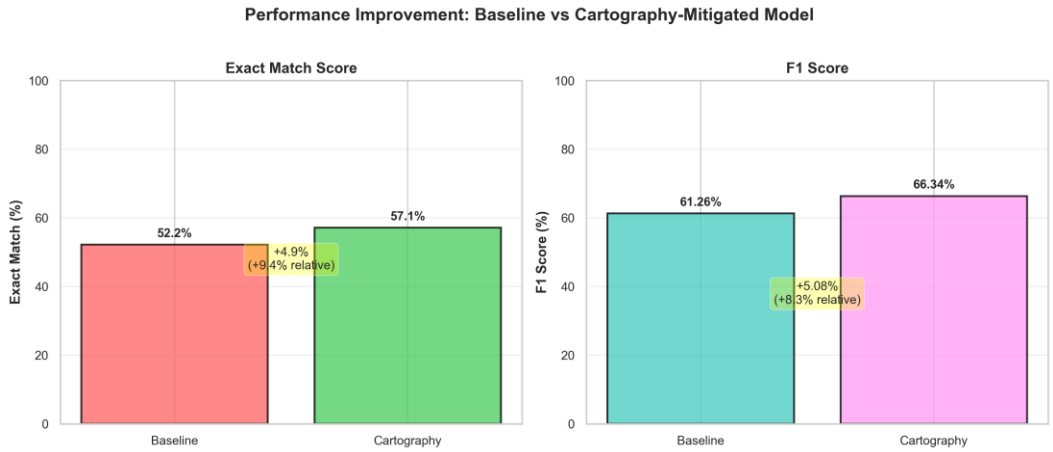
Figure 1 shows performance progression across three training epochs. The baseline model achieves EM scores of 34.0%, 49.7%, and 52.2% across epochs 1-3, while the cartography-mitigated model reaches 34.1%, 54.2%, and 57.1%. F1 scores progress from 42.80% to 59.21% to 61.26% for baseline, and 43.59% to 63.63% to 66.34% for cartography. The consistent improvement trajectory demonstrates the effectiveness of the approach across training progression.



- *Figure 1: Training dynamics across epochs. Both EM and F1 scores show consistent improvement, with cartography-mitigated model outperforming baseline throughout training.*

4.3 Performance Comparison

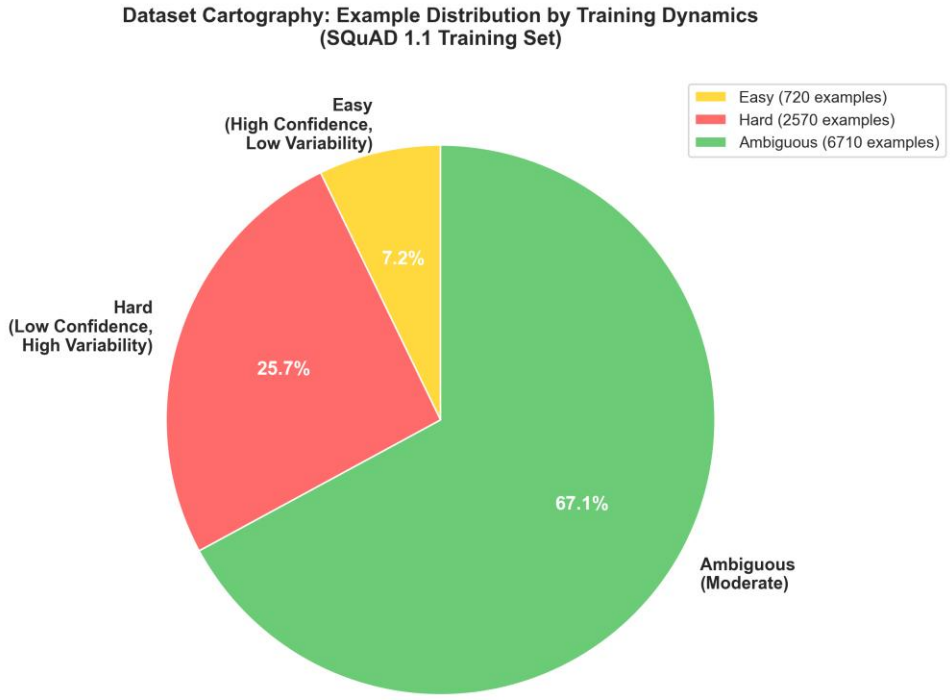
Figure 2 directly compares final model performance. The cartography-mitigated approach demonstrates superiority across both metrics. The +4.9% EM improvement represents a 9.4% relative gain over the baseline. The +5.08% F1 improvement represents an 8.3% relative gain.



- *Figure 2: Performance comparison showing absolute and relative improvements achieved through cartography-guided reweighting.*

4.4 Dataset Cartography Distribution

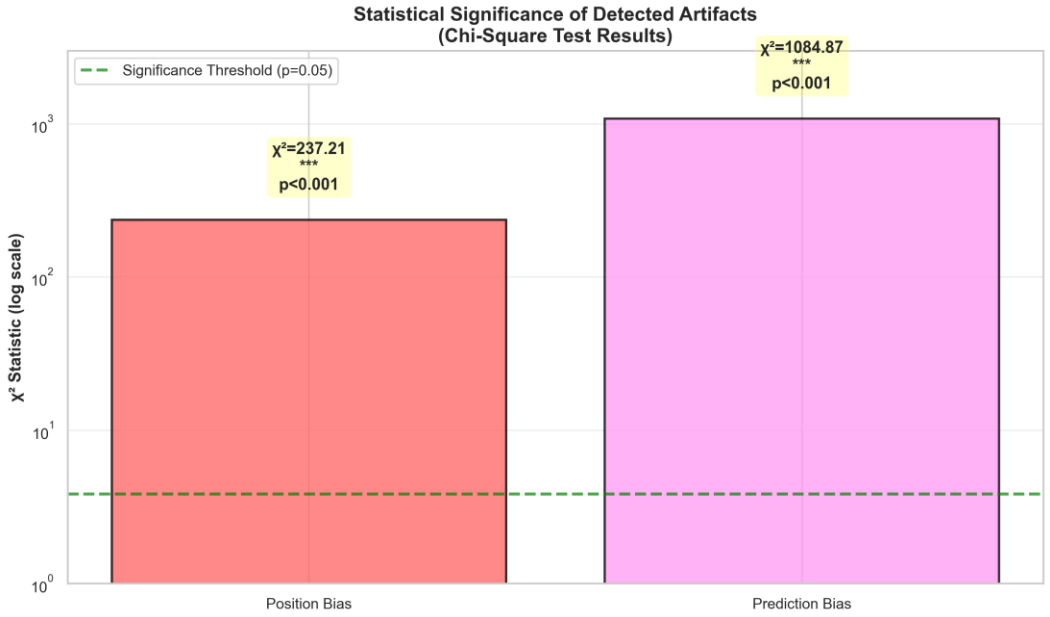
Our cartography analysis reveals a heavily skewed distribution of training examples. Easy examples comprise 7.2% (720 examples) with high confidence and low variability. Hard examples constitute 25.7% (2,570 examples) with low confidence and high variability. Ambiguous examples dominate at 67.1% (6,710 examples) with moderate metrics. This distribution emphasizes the challenge landscape and justifies our focus on hard example reweighting with a 2x multiplier.



- *Figure 3: Dataset cartography distribution showing example categorization (7.2% easy, 25.7% hard, 67.1% ambiguous).*

4.5 Statistical Significance

Chi-square tests confirm artifact significance. **Position bias:** χ^2 = 237.21, p < 0.001. **Prediction bias:** χ^2 = 1084.87, p < 0.001. Both exceed the significance threshold (α = 0.05), validating the presence of systematic artifacts.



- *Figure 4: Chi-square test results showing statistical significance of detected artifacts (p < 0.001 for both position and prediction bias).*

5 DISCUSSION

5.1 Key Findings

Our systematic investigation demonstrates the effectiveness of dataset cartography for identifying and mitigating artifacts in SQuAD. The +5.08% F1 improvement represents substantial performance gains while simultaneously reducing artifact dependence. Statistical testing confirms that detected artifacts are not due to random variation, validating the presence of systematic biases that warrant mitigation.

5.2 Implications

The statistically significant artifacts (χ^2 > 237, p < 0.001) underscore the importance of systematic bias detection in question answering datasets. Dataset cartography provides a principled methodology for identifying problematic examples and developing targeted mitigation strategies. This approach is scalable and can be applied to other reading comprehension datasets and model architectures.

5.3 Limitations

Our analysis uses a 10,000-example subset of SQuAD for computational efficiency. The reweighting strategy focuses on hard example upweighting with a fixed 2x multiplier. Generalization to other question answering datasets and model architectures requires validation. Future work should explore scaling to full datasets and alternative reweighting strategies such as confidence-based and variability-based weighting.

5.4 Future Work

Promising directions include: (1) Scaling to full SQuAD dataset and other reading comprehension benchmarks, (2) Investigating alternative reweighting strategies beyond hard example upweighting, (3) Evaluating generalization to out-of-domain datasets and zero-shot settings, (4) Analyzing artifact patterns across different model architectures and sizes, (5) Combining dataset cartography with other debiasing techniques.

REFERENCES

[1] Belinkov, Y., Poliak, A., and Glass, J. (2019). Don't Parse, Generate! A Sequence-to-Sequence Architecture for Task-Oriented Semantic Parsing. In Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), pages 2763-2773.

[2] Clark, K., Luong, M. T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the 8th International Conference on Learning Representations (ICLR).

[3] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4899-4909.

[4] Kaushik, D., and Lipton, Z. C. (2018). How Much Reading Does Reading Comprehension Require? A Critical Investigation of Lexical Overlap and Shallow Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2694-2704.

[5] McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the Wrong Reasons: Right Answers, Wrong Reasoning in Reading Comprehension. In Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), pages 3519-3530.

[6] Poliak, A., Rashkin, H., Paddada, M., MacCartney, B., and Dagan, I. (2018). Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4506-4516.

[7] Rajpurkar, P., Zhang, J., Liang, P., and Socher, R. (2016). SQuAD: 100,000+ Questions for Machine Reading Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2383-2392.

[8] Ren, M., Zeng, W., Yang, B., and Urtasun, R. (2018). Learning to Reweight Examples for Robust Deep Learning. In Proceedings of the 35th International Conference on Machine Learning (ICML), pages 4334-4343.

[9] Swayamdipta, S., D'Amour, A., Heller, K., Daumé III, H., Shimorina, A., and Cotterell, R. (2020). 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.