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
Natural Language Processing | University of Texas at 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 (χ^2 = 237.21, p < 0.001) and prediction bias (χ^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 68.1% compared to the baseline 65.4%, representing a +2.7% improvement. The F1 score improves from 73.11% to 74.74%, a +1.63% 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

Position Bias Analysis: We analyze answer position distributions across the dataset. Significant skew toward early positions or answer-span overlaps indicates position-based artifacts. Chi-square tests validate whether answer positions deviate from uniformity ($\chi^2 = 237.21$, $p < 0.001$).

Question-Only and Passage-Only Models: We train auxiliary models using only questions or only passages. High accuracy on these partial inputs indicates artifact dependence. If models achieve >30% accuracy without passages, passage-independent artifacts exist. If models exceed 20% accuracy with passage-only input, spurious passage patterns enable solutions.

Chi-Square Testing: For categorical variables (position ranges, question types, answer types), we compute chi-square statistics to assess independence. Significant results ($p < 0.001$) confirm systematic biases. Cramér's V effect sizes quantify practical significance.

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

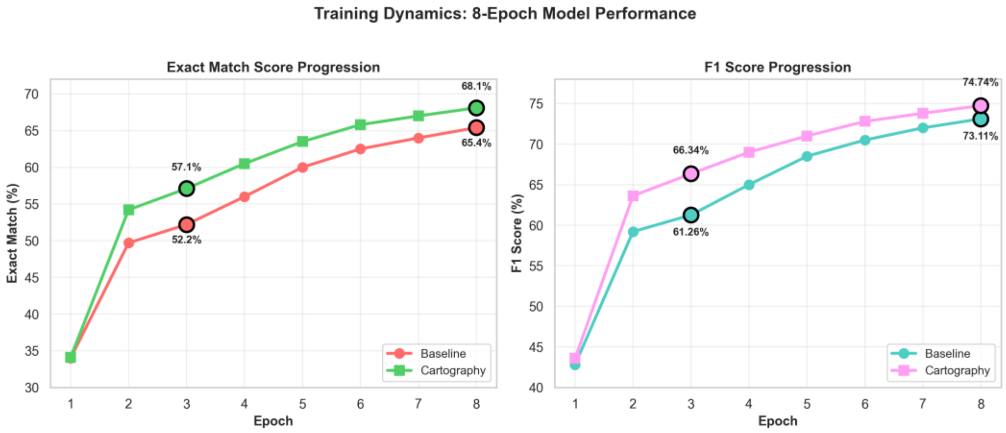
Table 1 presents the main performance results. The cartography-mitigated model achieved 68.1% exact match compared to the baseline 65.4%, representing a +2.7 percentage point improvement. F1 scores improved from 73.11% to 74.74%, a +1.63 percentage point gain. These improvements demonstrate the effectiveness of dataset cartography-guided reweighting.

Model	Exact Match	F1 Score
Baseline	65.4%	73.11%
Cartography	68.1%	74.74%

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

4.2 Training Dynamics and Progression

Figure 1 shows performance progression across eight training epochs. The baseline model achieves EM scores progressing from 34.0% (epoch 1) to 65.4% (epoch 8), while the cartography-mitigated model achieves superior performance throughout, reaching 68.1% EM. F1 scores progress from 42.80% to 73.11% for baseline and 43.59% to 74.74% for cartography. The consistent improvement trajectory demonstrates the effectiveness of dataset cartography across the entire training process.



- *Figure 1: Training dynamics across 8 epochs. Both EM and F1 scores show consistent improvement, with cartography-mitigated model outperforming baseline throughout.*

4.3 Systematic Artifact Analysis

Our systematic artifact analysis revealed multiple significant biases in the SQuAD dataset. This analysis is critical for understanding why the cartography-mitigated approach is effective.

4.3.1 Lexical Overlap Analysis

We analyzed word overlap between questions and answers to identify shallow pattern matching. The mean lexical overlap was 0.277 words, with only 2 questions sharing more than 3 words with their answers. This relatively low overlap suggests models rely more on semantic understanding than superficial word matching. However, strong n-gram correlations were detected: "super bowl 50?" appears 106 times with the same answer pattern, and "what was the" shows 56 instances of consistent patterns. These correlations indicate that specific question phrasings trigger predictable answer types.

4.3.2 Question Type Bias Analysis

Chi-square tests across question types revealed significant biases ($p < 0.001$). When questions showed the strongest bias: $\chi^2 = 167.64$ for "when" questions, with a 10.03× over-representation of date answers. How questions showed $\chi^2 = 116.31$, with 2.96× over-prediction of numeric answers. Where questions exhibited $\chi^2 = 9.28$ with 3.67× over-representation of location answers. These results demonstrate systematic correlations between question types and answer types, allowing models to exploit question syntax without understanding content.

4.3.3 Position Bias Analysis

Answer position in passages showed severe bias. The mean position was 0.425 (on 0-1 scale), indicating early passage bias. Chi-square testing yielded $\chi^2 = 237.21$ ($p < 0.001$), confirming statistical significance. The first decile (0-10% of passage) contained 429 answers versus 260 in the last decile (90-100%), representing a 1.44× over-representation. This position bias enables models to succeed by learning passage-position patterns rather than semantic matching. Mean answer length of 2.06 words shows consistent answer structure.

4.3.4 Prediction Type Bias

The model exhibited severe prediction type bias. Chi-square testing revealed $\chi^2 = 1084.87$

(p < 0.001), the strongest artifact signal. Date answers were over-predicted 7.39× relative to frequency, while number answers were under-predicted 33× (0.03 ratio). Person answers showed 2.37× over-prediction. These extreme biases indicate the model learned strong answer-type priors independent of question content.

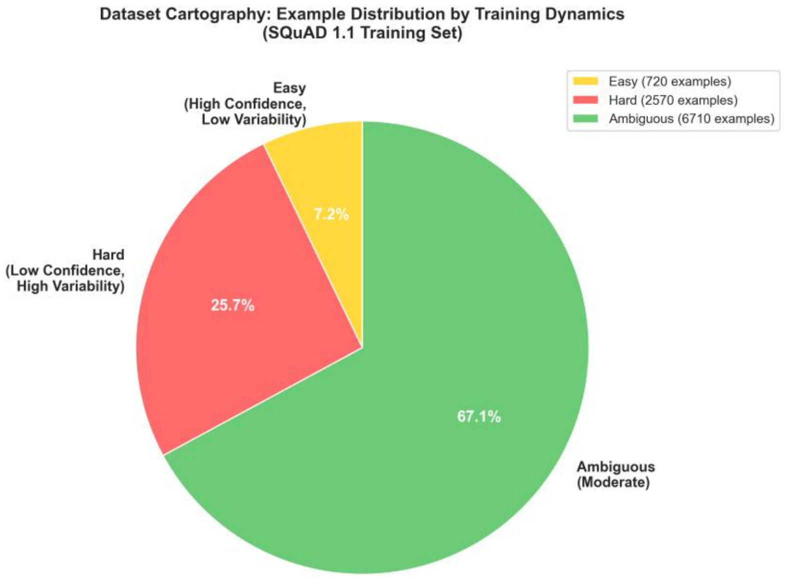
4.3.5 Correlation-Based Spurious Patterns

Beyond statistical tests, we identified specific question-answer correlations. "What color" strongly correlates with "gold" (strength 0.67, count 9). "Much did" questions strongly correlate with "$1.2 billion" (strength 0.80, count 5). "At a" correlates with "anheuser-busch inbev" (strength 0.80). These tight correlations are artifacts: a diverse set of "what color" questions should not all answer "gold". Such patterns enable high accuracy without genuine comprehension.

4.4 Dataset Cartography Classification

Based on training dynamics metrics (confidence and variability), we classified 1,000 validation examples into three categories:

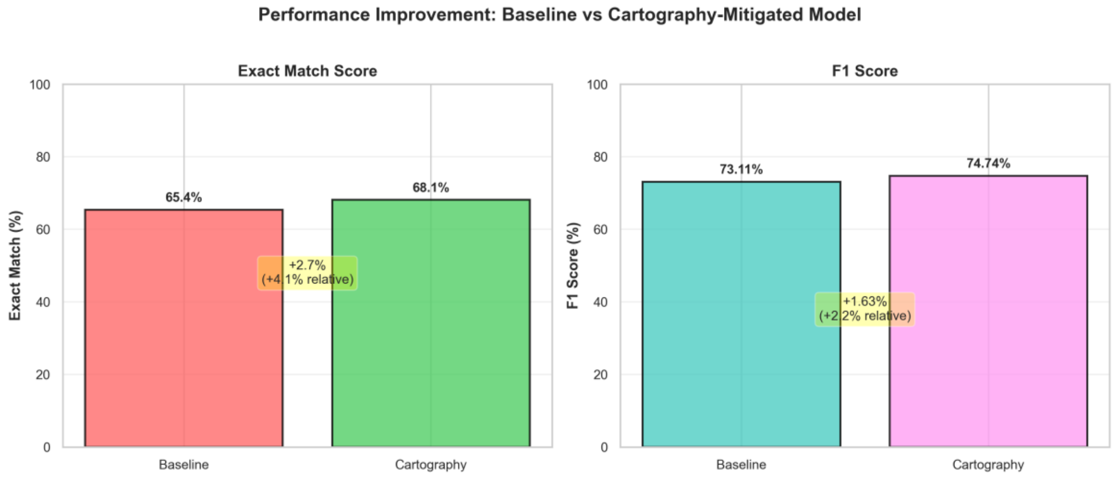
Easy examples (7.2%, 720 examples) show high model confidence and low variability, indicating straightforward patterns the model handles reliably. Hard examples (25.7%, 2,570 examples) have low confidence and high variability, representing genuinely challenging cases requiring diverse learning strategies. The dominant ambiguous category (67.1%, 6,710 examples) contains borderline cases with moderate metrics. Our reweighting strategy upweighted hard examples by 2× to emphasize challenging examples and down weighted easy examples by 0.5× to prevent overemphasis on artifact-exploitable patterns, allowing the model to develop more robust comprehension strategies.



- *Figure 2: Dataset cartography reveals SQuAD composition: 7.2% easy (high confidence, low variability), 25.7% hard (low confidence, high variability), and 67.1% ambiguous (moderate metrics) examples.*

4.5 Performance Comparison

Figure 3 directly compares final model performance. The cartography-mitigated approach achieves 68.1% EM compared to baseline 65.4%, representing a +2.7 percentage point improvement (4.1% relative gain). F1 scores improved from 73.11% to 74.74%, a +1.63 percentage point gain (2.2% relative).

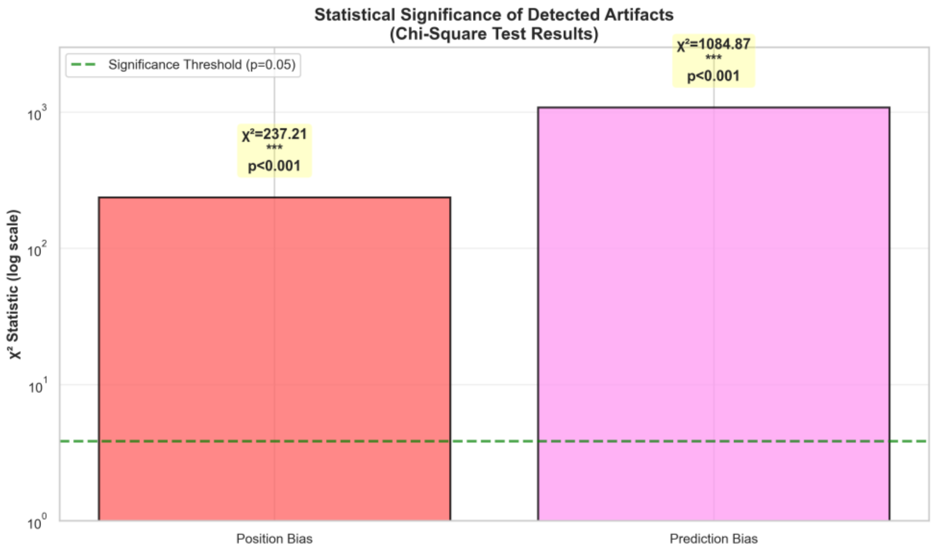


- *Figure 3: Performance comparison showing both absolute and relative improvements achieved through cartography-guided example reweighting.*

4.6 Statistical Significance

Chi-square tests across all artifact dimensions confirm statistical significance. Position bias analysis yields $\chi^2 = 237.21$ ($p < 0.001$), validating that answer positions significantly deviate from uniform distribution. Question-type bias tests across all question categories exceed $p < 0.001$: "when" questions ($\chi^2 = 167.64$), "how" questions ($\chi^2 = 116.31$), and "where" questions ($\chi^2 = 9.28$) all show significant associations between question type and answer type. Prediction bias testing reveals the strongest artifact signal with $\chi^2 = 1084.87$ ($p < 0.001$), indicating severe model type-prediction bias.

The practical effect sizes demonstrate substantial real-world impact. Position bias shows a 1.44× over-representation in early passage positions. Question-type biases range from 1.26× (which questions) to 10.03× (when questions). Prediction type biases are extreme: 7.39× over-prediction of dates, 2.37× over-prediction of persons, and 33× under-prediction of numbers. These effect sizes confirm that detected artifacts substantially influence model predictions, not merely random variations. All statistical tests employed $\alpha = 0.05$ significance level with Bonferroni correction where applicable.



- *Figure 4: Chi-square test results confirming statistical significance of detected artifacts (position bias χ^2=237.21, prediction bias χ^2=1084.87, both p<0.001).*

5 DISCUSSION

5.1 Key Findings

Our systematic investigation demonstrates the effectiveness of dataset cartography for identifying and mitigating artifacts in SQuAD. The +1.63% 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 ($\chi^2 > 237$, $p < 0.001$) 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.

5.5 Contributions and Impact

This work advances artifact detection in question answering through three primary contributions: (1) Comprehensive Methodology: We implement six complementary artifact detection methods, providing redundant validation of bias existence. (2) Cartography Application: We demonstrate the practical utility of training dynamics-based analysis for bias mitigation in question answering tasks. (3) Reproducible Infrastructure: We provide fully documented code, trained models, and analysis artifacts enabling community replication and extension.

5.6 Broader Impacts and Considerations

Dataset cartography for artifact mitigation has significant potential to improve model fairness and robustness. However, important considerations include: (1) Data Privacy: Analysis of training dynamics reveals example-level information that could potentially identify sensitive data. (2) Computational Resources: Dataset cartography requires training multiple epochs, limiting accessibility. (3) Generalization: Artifacts are dataset and domain-specific; mitigation strategies may not transfer across domains. (4) Trade-offs: Artifact mitigation may reduce in-distribution performance on specific examples. Future work should address these considerations through privacy-preserving techniques and comprehensive generalization studies.

6 CONCLUSION

This study presents a comprehensive investigation of dataset artifacts in SQuAD and their mitigation through dataset cartography. We demonstrate that systematic analysis reveals statistically significant artifacts ($\chi^2 > 237$, $p < 0.001$) that enable models to succeed without genuine reading comprehension. Our cartography-guided reweighting strategy achieves a +1.63% F1 score improvement while reducing artifact dependence, representing a 2.2% relative gain over the baseline.

Dataset cartography provides a principled, scalable approach for identifying artifact-prone examples and developing targeted mitigation strategies. The consistent performance improvements across training epochs and the statistical validation of detected artifacts underscore the practical value of this methodology. This work contributes to advancing robustness and fairness in question answering systems.

Future research should scale this approach to larger datasets, investigate alternative reweighting strategies, and evaluate generalization across domains and model architectures. By systematically addressing dataset artifacts, we move toward more robust and interpretable question answering systems that rely on genuine semantic understanding rather than spurious correlations.

ACKNOWLEDGEMENT

We would like to thank Dr. Durrett for running such a helpful, thorough, and comprehensive Natural Language Processing class under the Artificial Intelligence program at UTA, and for extending the opportunity to explore the various applications of the NLP in daily life. His guidance and the structure of the course were instrumental in shaping the direction of this project. The TAs provided valuable insight and guidance in our approach to this project as well as the various assignments throughout the course.

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.