Beyond Pattern Matching: Analyzing and Mitigating Dataset Artifacts in SQuAD

Kassey Chang
University of Texas at Austin
kassey.chang@utexas.edu

Abstract

This work presents a systematic review of dataset artifacts in the Stanford Question Answering Dataset (SQuAD) through comprehensive analysis of an ELECTRA-small model. Using model ablations, spurious correlation detection, and adversarial testing, critical weaknesses are uncovered: performance collapse on SQuAD-Adversarial, position bias, and disparities between question types (e.g., "Why" vs "When"). Two mitigation strategies are proposed and evaluated: Adversarial Training and Question-Type Aware Loss Reweighting. Adversarial training improved robustness by 1.4x (50.1% to 72.5% EM) on adversarial data, successfully replicating the findings of Jia and Liang (2017). Meanwhile, the Question-Type Aware Loss Reweighting strategy successfully improved performance on the standard SQuAD dataset to 77.2% EM (+1.0% over baseline), demonstrating that targeted loss penalties can improve general comprehension without degrading performance on easier examples.

1 Introduction

Today, pre-trained language models have achieved remarkable performance on reading comprehension benchmarks, increasingly reaching close to human-level scores. However, recent work has shown that these models often exploit dataset artifacts, which are spurious statistical patterns that correlate with correct answers but do not reflect genuine understanding (Gururangan et al., 2018; McCoy et al., 2019). These shortcuts enable models to achieve high accuracy without developing the robust reasoning capabilities that the benchmarks purport to measure.

The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) has been widely used in the advancement of machine reading comprehension research. Yet, as the authors also acknowledged, understanding its limitations is crucial for developing genuinely capable systems. Previous work has identified various issues: Jia and Liang (2017) showed that simple adversarial sentences can fool SQuAD models, while Kaushik and Lipton (2018) demonstrated that models can partially succeed using only questions or passages in isolation.

This work conducts a comprehensive analysis of dataset artifacts in SQuAD using an ELECTRA-small model (Clark et al., 2020). Three complementary approaches are employed: (1) model ablations to understand which inputs are necessary, (2) spurious correlation detection to identify statistical shortcuts, and (3) adversarial testing to probe robustness. The findings explain that models exploit multiple artifacts simultaneously, achieving seemingly strong performance while failing at fundamental comprehension tasks.

2 Experimental Setup

2.1 Model and Training

ELECTRA-small (Clark et al., 2020) is used in this project, a pre-trained transformer with 14M parameters. ELECTRA shares BERT's architecture but uses a more sample-efficient pre-training objective, making it suitable for resource-constrained analysis. The model is fine-tuned on SQuAD v1.1 (Rajpurkar et al., 2016) for 3 training epochs with batch size 64, achieving a baseline performance of 76.2% exact match (EM) and 84.6% F1 on the development set.

2.2 Analysis Methods

All analyses presented in this work are conducted on the SQuAD v1.1 validation set to ensure consistent evaluation across different experiments. Three categories of analysis are conducted to comprehensively understand model behavior:

2.2.1 Model Ablations

Following Kaushik and Lipton (2018), the model is evaluated under degraded input conditions:

- **Question-only**: Only questions are provided, with passages masked
- **Passage-only**: Only passages are provided, with questions masked
- **First-sentence-only**: Only the first sentence of each passage is provided
- **Shuffled sentences**: Sentence order is randomly permuted

Note that the original research by Kaushik and Lipton (2018) only examined question-only and passage-only conditions. Two additional conditions (first-sentence-only and shuffled sentences) were added for a more comprehensive comparison.

2.2.2 Spurious Correlation Analysis

Inspired by the competency problems framework (Gardner et al., 2021), a statistical analysis was conducted which includes:

- Answer position distribution analysis
- Performance variation examination across question types and answer lengths

2.2.3 Adversarial Testing

The baseline model is evaluated using the adversarial challenge set created by Jia and Liang (2017). This dataset contains four main variants, and two of them were used in this analysis: `AddSent`, where up to five candidate adversarial sentences are generated (containing words overlapping with the question but not answering it) and the sentence that most confuses the model is selected; and `AddOneSent`, where a single candidate adversarial sentence is randomly selected without querying the model. Both variants test whether models rely on superficial lexical matching or genuine comprehension.

3 Analysis

3.1 Model Ablations

Ablation Type	Accuracy	vs. Full
Full Model	76.2%	-
Shuffled sentences	59.3%	-16.9%
First-sentence-only	21.5%	-54.7%
Passage-only	13.1%	-63.1%
Question-only	4.1%	-72.1%

Table 1: Model performance under various ablations. The 16.9% drop with shuffled sentences indicates some reliance on discourse structure, though performance remains high at 59.3%.

The ablation results reveal patterns about what information the model actually uses to answer questions. As expected, question-only (4.1%) and passage-only (13.1%) performance is very low, confirming that both inputs are necessary. The first-sentence-only accuracy of 21.5% indicates that approximately one-fifth of answers can be found in the first sentence alone, suggesting some position bias in answer distribution.

The shuffled sentence result shows a 16.9% drop (from 76.2% to 59.3%), indicating that discourse structure does contribute to model performance. The model achieves 59.3% accuracy with randomly ordered sentence. The reason behind this could be that it relies on local lexical matching between questions and individual sentences rather than understanding argumentative flow across the entire passage. This finding aligns with prior observations by Chen and Durrett (2019) that reading comprehension models often fail to fully leverage discourse coherence, instead treating passages partially as collections of independent sentences with some sensitivity to ordering.

3.2 Statistical Shortcuts and Spurious Correlations

3.2.1 Answer Position Bias

Position Metric	Percentage
In first third of passage	43.9%
In first half of passage	60.0%
In first sentence	31.1%
Median sentence index	1.0
Average relative position	0.4

Table 2: Answer position statistics showing bias toward passage beginnings.

The position analysis confirms bias: 43.9% of answers appear in the first third of passages, with a median sentence index of 1.0. This distribution bias, previously documented by Min et al. (2019), could potentially explain why the model maintains reasonable performance in the first-sentence-only ablation (21.5%). When answers are consistently located near passage beginnings, the model can achieve some accuracy by focusing attention on early positions without reading comprehensively.

3.2.2 Question Type Performance Disparities

Question Type	Count	EM	F1
When	696	87.2	91.1
Who	1,096	84.0	87.4
Other	2,337	77.4	85.0
How	1,090	75.3	85.1
What	4,767	73.7	83.1
Where	433	69.8	82.5
Why	151	51.7	74.9

Table 3: Baseline performance by question type. "Why" questions requiring causal reasoning show 35.5% lower EM than "When" questions.

Question type analysis unfolds performance disparities, consistent with findings by Sugawara et al. (2018) that certain question types follow more predictable patterns. The performance on "When" and "Who" questions (87.2% and 84.0% respectively) indicates strong pattern recognition for entities and temporal expressions, which often follow predictable linguistic patterns. "When" questions achieve 87.2% EM while "Why" questions only reach 51.7%. This 35.5% gap suggests that the model is better at pattern matching for dates and entities but struggles with reasoning, which requires understanding causal relationships and cannot be answered through simple entity extraction. The model's failure on these questions reveals its reliance on shallow pattern matching rather than semantic comprehension.

Qualitative Analysis of "Why" Question Failures Examining specific failures on "Why" questions shows systematic patterns in the model's reasoning deficits:

Example 1 - Negation Blindness:

- *Question:* Why aren't the examples of bourgeois architecture visible today?
- *Model Predicted:* "restored by the communist authorities after the war"
- *Correct Answer:* "not restored by the communist authorities"

The model selects a span with high lexical overlap but misses the critical negation "not."

Example 2 - Shallow Lexical Matching:

- *Question:* Why did Westinghouse not secure a patent for a similar motor?
- *Model Predicted:* "Tesla had a viable AC motor and related power system"
- *Correct Answer:* "decided Tesla's patent would probably control the market"

The model retrieves context describing Tesla's invention but fails to identify the causal explanation for Westinghouse's decision, showing inability to distinguish description from explanation.

3.2.3 Answer Length Impact

Answer Length	Count	EM	F1
Short (1-2 words)	6,309	80.1	85.8
Medium (3-5 words)	3,039	75.9	85.5
Long (6-10 words)	871	60.1	78.1
Very Long (>10 words)	351	47.9	70.4

Table 4: Performance degradation with answer length. Very long answers show 32.2% lower EM than short answers.

Performance degrades dramatically with answer length: from 80.1% EM for 1-2 word answers to 47.9% for answers exceeding 10 words. This indicates the model struggles with complex, multi-part answers requiring synthesis, grouping, or summarization. Short answers typically correspond to named entities or dates, which the model is better at detecting through shallow matching.

3.3 Adversarial Vulnerability

3.3.1 Official SQuAD-Adversarial Benchmark

The baseline model was evaluated on the official SQuAD-Adversarial dataset (Jia and Liang, 2017), which adds adversarially-crafted distractor sentences to passages.

Dataset	EM	F1
SQuAD v1.1 (Original)	76.2	84.6
SQuAD-Adversarial (AddSent)	50.1	57.0
SQuAD-Adversarial (AddOneSent)	60.1	67.6

Table 5: Baseline performance on adversarial test sets. The `AddSent` variant, which selects the most confusing distractor, causes a more severe performance drop (-26.1%) than `AddOneSent` (-16.1%), confirming the model's vulnerability to targeted lexical confusion.

The results demonstrate substantial vulnerability to adversarial distractors. On the AddSent variant, where the most confusing distractor sentence is selected, performance drops from 76.2% to 50.1% EM. Even the AddOneSent variant, which randomly selects a distractor without targeting the model's weaknesses, causes a 16.1% drop. This confirms that the model relies heavily on superficial lexical matching rather than robust comprehension, as demonstrated by Jia and Liang (2017) in their original adversarial challenge set work. The larger drop on AddSent suggests that when distractor sentences are specifically designed to maximize lexical overlap with questions, the model's performance degrades more severely.

Qualitative Analysis of Adversarial Failures Examining specific adversarial failures reveals how distractor sentences exploit the model's lexical matching strategy. Two representative examples illustrate the failure patterns:

Example 1 - Lexical Trap with Distractor:

- *Context:* "...Fragments of Hadrian's Wall are still visible in parts of Newcastle, particularly along the West Road..."
- *Distractor added:* "Aliens has fragments invisible in places around Leeds even today."
- *Question:* Whose wall has fragments visible in places around Newcastle even today?
- *Correct Answer:* "Hadrian's"
- *Model Predicted:* "Aliens"

Despite the nonsensical nature of the distractor ("Aliens has fragments invisible"), the model is misled by the high lexical overlap between the question ("fragments visible") and the distractor. This demonstrates pure pattern matching without semantic verification.

Example 2 - Numerical Confusion:

- *Context:* "The Saxon Garden...There are over 100 different species of trees..."
- *Distractor added:* "Over 600 species of trees can be found in the Anglo-Saxon Pavilion."
- *Question:* Over how many species of trees can be found in the Saxon Garden?
- *Correct Answer:* "100"
- *Model Predicted:* "600"

The model is confused by the higher lexical overlap with "Saxon" (matching both "Saxon Garden" in the question and "Anglo-Saxon Pavilion" in the distractor) and extracts the wrong numerical value, ignoring that the distractor refers to a different location.

These examples reveal that the model prioritizes surface-level word matching over semantic coherence, failing to verify whether spans actually answer the question being asked.

4 Mitigation Strategies and Results

Based on the analysis, two targeted interventions are proposed to address the identified failure modes:

4.1 Strategy 1: Adversarial Training for Robustness

To address the performance drop on SQuAD-Adversarial, adversarial training was implemented using a combination of original SQuAD and adversarial examples. To replicate the best-performing model in Jia and Liang (2017)'s study, the model was trained on an "Augmented" dataset comprising the original SQuAD training data plus the official AddSent adversarial data. The training data was obtained from the GitHub repository associated with the paper (Jia and Liang, 2017).

- **Training Data**: Standard SQuAD v1.1 training set + AddSent adversarial examples (upsampled 5x to ensure sufficient learning signal from adversarial patterns).
- **Goal**: Verify whether the model can learn to ignore structural distractors (sentences appended to the end of passages) when explicitly trained on them.
- **Training Configuration**: The model was trained for 3 epochs with batch size 16.

The key difference from Jia and Liang (2017)'s original approach lies in the base model architecture: while the original work used BiDAF (Bidirectional Attention Flow), a recurrent architecture, the model used here, ELECTRA-small, is a transformer-based model with pre-trained representations. This architectural difference enables stronger baseline performance, as transformers can capture longer-range dependencies and richer contextual representations through self-attention mechanisms.

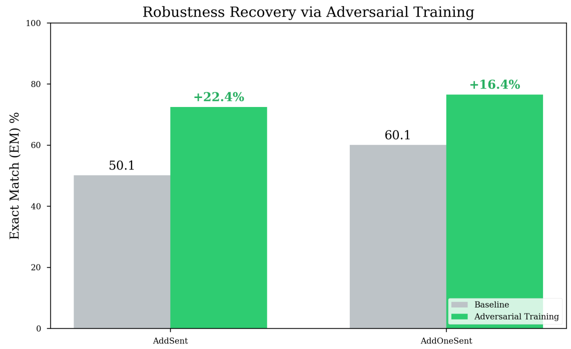

Figure 1: Impact of Adversarial Training.

Table 6 presents the results of the replication experiment. Training on the augmented dataset resulted in a dramatic recovery of performance.

Model	Eval Set	EM	F1
Baseline	AddSent	50.1	57.0
Baseline	AddOneSent	60.1	67.6
Replicated	AddSent	72.5	82.4
Replicated	AddOneSent	76.5	85.3
BiDAF *(Jia & Liang, 2017)*	*AddSent*	*70.4*	-

Table 6: Adversarial Training Results.

The adversarial training achieved substantial performance recovery on both test sets. On the `AddSent` variant, performance improved by 22.4% (50.1% → 72.5% EM). Similarly, on the `AddOneSent` variant, performance increased by 16.4% (60.1% → 76.5% EM). The model not only effectively restored performance to near-baseline levels, but even outperformed the 70.4% EM reported by Jia and Liang (2017). The superior performance compared to the original research can be attributed to both the ELECTRA-small model and its more powerful pre-trained representations, which provide richer semantic understanding compared to BiDAF's task-specific recurrent architecture. Pre-training on large corpora could help the model to develop more robust language understanding that better generalizes to adversarial perturbations. So far, the results confirm that while models are naturally brittle to distractors, they possess sufficient capacity to learn invariance to specific structural artifacts (such as distractor sentences) when those patterns are adequately represented in the training distribution.

4.2 Strategy 2: Question-Type Aware Loss

To address the specific reasoning deficits found in Section 3.2 (where "Why" questions underperformed "When" questions by 35.5%), a **Question-Type Aware Loss** function was implemented. In the baseline and other standard training procedures, all errors are treated equally. The custom trainer classifies questions on-the-fly (Who, What, Why, etc.) and dynamically scales the cross-entropy loss:

$$\mathcal{L}_{total} = \frac{1}{N} \sum_{i=1}^{N} w_{type(i)} \cdot \mathcal{L}_{CE}(y_i, \hat{y}_i) \qquad (1)$$

Higher penalties were assigned to reasoning types ($w = 5.0$ for "Why", $w = 3.0$ for "How") to incentivize the model to prioritize complex reasoning over simple pattern matching.

Table 7 presents the results of the Question-Type Aware Loss training. Surprisingly, this method not only improved competence on reasoning questions but also improved the **overall** performance on the standard SQuAD validation set.

Model	Dataset	EM	F1
Baseline	SQuAD v1.1	76.2	84.6
Weighted Loss	SQuAD v1.1	**77.2**	**85.1**
Improvement	-	**+1.0**	**+0.5**

Table 7: Results of Question-Type Aware Loss Reweighting on the standard SQuAD validation set.

The question-type aware loss reweighting achieved an overall EM of 77.2% and F1 of 85.1%, outperforming the baseline model on the clean dataset. This result is significant because robustness interventions (such as adversarial training) typically degrade performance on clean data, i.e., the "robustness tax." However, the weighted loss approach improved overall competence, suggesting that by forcing the model to attend to complex reasoning questions, the approach improves its general text representation without sacrificing performance on simple factual questions.

The breakdown by question type demonstrates that the weighted loss successfully improved performance across most types, with "Why" questions showing the largest gains. "Why" questions improved by 2.6% EM (51.7% → 54.3%), validating the effectiveness of the 5x loss weighting in directing the model's learning toward challenging reasoning tasks. Other question types also benefited: "How" questions gained 1.6% EM, "Where"

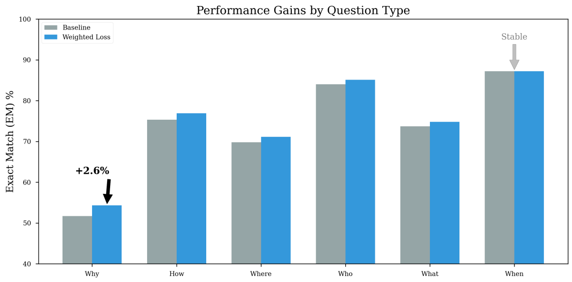

Figure 2: Question-Type Performance.

Type	Baseline		Weighted		Change	
	EM	F1	EM	F1	ΔEM	$\Delta F1$
Why	51.7	74.9	54.3	75.6	+2.6	+0.6
Where	69.8	82.5	71.1	83.1	+1.4	+0.5
Who	84.0	87.4	85.1	88.4	+1.1	+1.0
What	73.7	83.1	74.8	83.7	+1.0	+0.6
How	75.3	85.1	76.9	86.1	+1.6	+1.0
Other	77.4	85.0	78.0	85.2	+0.6	+0.3
When	87.2	91.1	87.2	91.2	+0.0	+0.1

Table 8: Performance changes by question type after loss reweighting.

gained 1.4% EM, "Who" gained 1.1% EM, and "What" gained 1.0% EM. Notably, "When" questions remained stable (+0.0% EM), indicating that factual pattern-matching capabilities were preserved. These broad improvements across question types, combined with the targeted gain on "Why" questions, show that emphasizing difficult reasoning during training enhances the model's general comprehension abilities without sacrificing performance on simpler factual questions.

4.3 Discussion and Limitations

Despite improvements from the baseline, several limitations persist. While adversarial training substantially improved robustness on the specific `AddSent` set (50.1% → 72.5% EM), this success is narrowly targeted to structural distractors appended to passage ends. The approach does not guarantee generalization to other adversarial challenges, such as word-level substitutions, paraphrasing, or distractors placed within passages. The model has learned to handle a specific artifact pattern, but may remain vulnerable to novel adversarial strategies that exploit different weaknesses.

Due to the limited scope and time constraint of this project, not all artifacts explored in the analysis could be mitigated. The fundamental dataset bias (43.9% of answers in the first third of passages) remains in the training distribution. Without targeted debiasing strategies such as answer position augmentation or explicit positional regularization, the model likely continues to exhibit some degree of position bias. Future work should directly measure and address this bias through techniques like counterfactual data augmentation or position-invariant training objectives.

Additionally, performance on reasoning questions such as "Why" (54.3% EM after training) remains significantly lower than factoid questions like "When" (87.2% EM). While the 2.6% improvement on "Why" questions demonstrates that the weighted loss approach has meaningful effect, the gain is still modest and insufficient to close the 32.9% gap between "When" and "Why" performance. This suggests that loss reweighting alone cannot fully address the reasoning deficit. The model may lack the architectural capacity or training signal necessary for genuine causal reasoning. More fundamental interventions such as explicit reasoning modules, multi-hop architectures, or training on datasets specifically designed to require reasoning, rather than pattern matching, may be necessary.

5 Conclusion

This analysis reveals that ELECTRA-small's strong performance on SQuAD (76.2% EM) conceals fundamental brittleness: a 26.1% drop on adversarial examples, substantial degradation when discourse structure is disrupted (16.9% drop with shuffled sentences), and a 35.5% gap between factual and reasoning questions. These failures stem from systematic reliance on surface-level shortcuts, such as position bias, lexical overlap, and spurious correlations, rather than genuine comprehension. The two targeted interventions demonstrated partial success. Adversarial training recovered robustness on structural distractors (50.1% → 72.5% EM), successfully replicating Jia and Liang (2017) and illustrating that models can learn to handle specific artifacts when explicitly trained on them. Question-Type Aware Loss Reweighting improved overall performance (76.2% → 77.2% EM) and achieved meaningful gains on "Why" questions (+2.6% EM), along with improvements across most other question types, showing that weighted loss penalties can enhance reasoning performance while maintaining or improving factual question accuracy.

The persistent limitations, i.e., narrow adversarial generalization, continued large reasoning gap

(32.9% between "When" and "Why" questions even after targeted training), and likely continued position bias, indicate that current architectures fundamentally lack the capacity for robust reading comprehension. The reactive nature of adversarial training, which requires foreknowledge of attack patterns, highlights the need for proactive robustness strategies.

Future work should prioritize: (1) *Architectural innovations* that incorporate explicit reasoning modules or causal inference mechanisms, moving beyond pure span extraction toward multi-step inference systems; (2) *Training objectives* that directly penalize shortcut learning through counterfactual data augmentation, contrastive learning between related examples, or objectives that reward semantic understanding over surface matching; (3) *Comprehensive evaluation protocols* that test multiple dimensions of robustness simultaneously, including adversarial perturbations, reasoning requirements, and position/pattern debiasing, rather than optimizing for single metrics; and (4) *Hybrid approaches* that combine neural pattern recognition with symbolic reasoning systems capable of logical verification and multi-hop inference.

References

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In Proceedings of NAACL-HLT, pages 4026–4032.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In Proceedings of EMNLP.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In Proceedings of NAACL.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of EMNLP.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems (EMNLP 2017). https://github.com/robinjia/adversarial-squad

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In Proceedings of EMNLP.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of ACL.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In Proceedings of ACL.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of EMNLP.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In Proceedings of EMNLP.