Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/GC02.20240100
Published: December 01, 2024

Machine Learning for Legal Contract QA: Analyzing and Enhancing ELECTRA with Adversarial Triggers

Amir Barakati
Associate Professor
Department of Mechanical and Aerospace Engineering
George Washington University
Washington DC 20052, USA
amir.barakati@gwu.edu

Arash Mehdizadeh
Assistant Professor
Department of Electrical and Electronics Engineering
College of Engineering, Australian University
West Mishref, Safat 13015, Kuwait
a.mehdizadeh@au.edu.kw

Baktanoosh K. Nakhjavani
Lead Engineer
FATA Automation
Department of Engineering
Auburn Hills MI 48326, USA
baktanoosh@gmail.com

Abstract

Natural Language Processing (NLP) models have become essential for handling tasks such as Question-Answering (QA). This is particularly true in domains that require high accuracy, like legal document analysis. This study evaluates the performance of the ELECTRA model in QA tasks by testing it on both Wikipedia-based and legal contract datasets, including the Contract Understanding Atticus Dataset (CUAD). An adversarial attack known as a universal adversarial trigger was then introduced. This trigger is designed to test the robustness of the ELECTRA model by introducing small perturbations that aim to confuse the model into making incorrect predictions. Our results show that the adversarial trigger reduces the model's performance by more than 10% on the well-known SQuAD dataset. This drop in accuracy highlights the vulnerability of the model to adversarial manipulation. On the legal contract dataset, specifically CUAD, we achieved a performance improvement of over 12% after fine-tuning the model with targeted optimizations. These optimizations included adjusting the model's hyperparameters and modifying the dataset to reduce ambiguity in question phrasing. These findings indicate that while ELECTRA performs well on general datasets, it struggles with the intricacies of legal texts. However, with specific optimizations and adjustments, the model can become more effective in legal document analysis.

Keywords

Question-Answering, Natural Language Processing, ELECTRA, Legal Contracts, Adversarial Triggers

1. Introduction

Extracting information from a given context is essential in many applications. With the rapid growth of data, this task has become increasingly difficult. Advances in Natural Language Processing (NLP) have enabled the development of state-of-the-art models to search documents for relevant answers (Calijorne Soares & Parreiras, 2020). Question-Answering (QA) systems are one such solution. However, achieving accurate predictions requires proper training of these models for different types of QA tasks (Nassiri & Akhloufi, 2023).

An exciting field of study entails enhancing the resilience of these NLP models against adversarial attacks. Such attacks can lead the models to produce errors, resulting, for instance, in bypassing a spam email filter by a spammer (Biggio et al., 2013; Goyal et al., 2023; Jia & Liang, 2017; Kuchipudi et al., 2020; Zhang et al., 2020). Triggers are a form of adversarial perturbation added to the input context to trick the model into producing a specified wrong prediction (Moosavi-Dezfooli et al., 2017; Shao et al., 2022). (Wallace et al., 2019) used a gradient-guided search over tokens to select and design triggers. The tokens in the trigger sequence are updated in each iteration to increase the likelihood of the target prediction for batches of examples (Wallace et al., 2019).

This study focuses on legal contract documents, a challenging task, even for the state-of-the-art models, due to the lack of appropriate datasets, expensive preparation, and different structure of documents. The preparation or review of commercial legal contracts is a costly and time-consuming process. AI-based models can reduce the cost of contract review focused on critical clauses with higher precision. Recent research has proven that using AI software can reduce costly human contract reviews by 60% (Hassan et al., 2021; Hendrycks et al., 2021). Although this method still cannot be used for Judicial judgment, there are many areas where this approach can be useful or increase precision.

In this study, several Wikipedia- and legal contract-based QA datasets are used to train and test the ELECTRA model. The model is fine-tuned using an experimental approach. The performance of the pretrained model is then evaluated and hypotheses are proposed to possibly improve the accuracy of the model on the question-answering task related to important clauses of legal documents.

1.1 Objectives

The objectives of this research are as follows. I) to evaluate the performance of the ELECTRA model in answering questions from generic and then more complex contexts. II) to introduce and test a universal adversarial trigger on NLP models, particularly ELECTRA, to assess vulnerability and robustness in complex contexts such as legal documents. III) To identify potential improvements in QA accuracy through fine-tuning and dataset modifications.

2. Literature Review

QA systems have become an essential tool within the broader field of NLP. These systems are designed to extract precise answers from large volumes of text, making them invaluable in fields like customer service, education, and more recently, legal document analysis (Calijorne Soares & Parreiras, 2020). QA models typically rely on pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers), which use contextual embeddings to understand text (Devlin et al., 2019; Kaliyar, 2020). These advancements have allowed models to accurately interpret and respond to a variety of questions, but challenges remain, especially when dealing with complex, domain-specific texts such as legal contracts (Hendrycks et al., 2021).

Legal documents present unique difficulties for NLP models due to their specialized language and structure. The interpretation of legal texts often requires deep domain knowledge and the ability to understand lengthy, intricately worded clauses. Standard QA models, which perform well on simpler texts like Wikipedia articles, often struggle when faced with this kind of complexity. Legal documents frequently contain jargon, ambiguous phrasing, and terms of art, which complicate the QA task (Ganguly et al., 2023). This has driven the development of domain-specific datasets, such as the Contract Understanding Atticus Dataset (CUAD) dataset, which provides a large corpus of annotated legal contracts for QA tasks (Hendrycks et al., 2021).

Transformer-based models like BERT, GPT, and ELECTRA have significantly improved the ability of QA systems to handle complex texts. BERT, for instance, introduced the concept of bi-directional learning, which allowed models to understand context by analyzing words both before and after a target word in a sentence (Devlin et al., 2019; Kaliyar, 2020). This innovation led to dramatic improvements in tasks such as reading comprehension and question-

answering. However, despite BERT's success, it has limitations, particularly in terms of computational efficiency and its ability to handle adversarial attacks.

ELECTRA was developed to address some of these limitations. It improves on BERT by using a generator-discriminator architecture, where one model generates synthetic inputs and another detects whether those inputs are real or fake (Clark et al., 2019). This approach requires less computational power while delivering comparable or even superior results to BERT on a range of tasks, including QA. ELECTRA's use of discriminative learning helps it excel in distinguishing between nuanced legal terms, making it a promising candidate for legal document analysis.

One of the major concerns in QA systems is their vulnerability to adversarial attacks. Adversarial attacks involve making small, often imperceptible changes to input data that cause models to produce incorrect answers (Biggio et al., 2013; Goyal et al., 2023; Moosavi-Dezfooli et al., 2017; Wallace et al., 2019). These attacks have been widely studied in the field of image recognition, but they also pose a significant challenge to NLP models. (Jia & Liang, 2017) were among the first to explore adversarial attacks in QA systems, showing that even well-performing models like BERT could be fooled by carefully crafted questions or context modifications.

The concept of "universal adversarial triggers" introduced by (Wallace et al., 2019) has gained attention due to its effectiveness in disrupting QA models. These triggers are sequences of words that can be inserted into any context to cause a model to fail, regardless of the actual content of the text. For example, Wallace's work demonstrated how inserting nonsensical phrases could dramatically reduce the accuracy of QA models, especially on datasets like The Stanford Question Answering Dataset (SQuAD). Legal documents, with their dense and intricate language, are particularly vulnerable to such adversarial perturbations, making it essential to develop models that can resist these attacks.

To effectively train QA models for the legal domain, high-quality, domain-specific datasets are required. SQuAD has been widely used for training and evaluating QA models (Rajpurkar et al., 2016). SQuAD contains questions and answers based on Wikipedia articles and is an excellent benchmark for general QA tasks. However, it does not capture the complexity and specificity of legal texts.

To address this gap, (Hendrycks et al., 2021) introduced the CUAD dataset, a specialized dataset designed for legal contract review. CUAD includes over 13,000 labels for critical clauses in a wide range of contract types, including non-disclosure agreements, purchase orders, and employment contracts. The dataset's complexity reflects real-world legal challenges, such as identifying clauses that require special attention from lawyers. CUAD's size and scope make it a valuable resource for training models to handle legal documents, but it also highlights the limitations of current QA systems in this domain.

A recent refinement of the CUAD dataset is the Filtered CUAD, which removes older contracts and reduces the number of clause categories from 41 to 12 (Apostolo, 2022). This modified dataset focuses on the most critical legal terms and improves the model's ability to analyze modern contracts efficiently. The filtered version was developed to reduce processing time and confusion caused by outdated legal terminology, further enhancing the model's performance.

Another variation is the CUAD QA dataset, where questions are simplified to match the labels of legal clauses more directly (Hao, 2022). This dataset aims to improve the model's focus on the specific legal language rather than the full contract context. The simplified format reduces ambiguity in questions, which is particularly important for QA models that often struggle with the complex phrasing found in legal texts.

3. Methods

This section outlines the steps to evaluate the performance of ELECTRA across various QA datasets, including general-purpose and domain-specific legal datasets. The ELECTRA model was used as the primary framework for training, testing, and validating various QA datasets in this study. ELECTRA is a transformer-based model designed to distinguish between real and replaced tokens, improving both training efficiency and accuracy compared to previous models like BERT (Clark et al., 2019). ELECTRA employs a two-step process: first, a generator creates "fake" tokens by replacing certain words in the input text; then, a discriminator predicts whether each word in the text is real or

generated.

The ELECTRA-small variant was chosen for this research due to its balance between computational efficiency and performance. This model can learn contextual embeddings with less computational cost than its larger counterparts, making it ideal for extensive experimentation across multiple datasets.

The first step in the experiment was to train the model on SQuAD and then test it on the Adversarial QA dataset. SQuAD serves as a benchmark for reading comprehension tasks, and the model's performance on this dataset provided the baseline evaluation (Rajpurkar et al., 2016).

After establishing a baseline, the model was tested on the Adversarial QA dataset, which uses the same articles as SQuAD but poses questions specifically designed to confuse the model (Jia & Liang, 2017). This was a critical test for assessing the model's resilience against adversarial perturbations. The hyperparameters were fine-tuned based on results from these initial tests to optimize performance gains.

A universal adversarial trigger was designed to investigate how small, targeted modifications to the input could disrupt model performance. This method was based on the approach proposed by (Wallace et al., 2019), where a sequence of tokens is added to the input context to confuse the model. The trigger was tested on the SQuAD dataset to assess performance. The design of this adversarial trigger was relatively simple compared to gradient-guided methods but still proved effective in reducing the model's prediction accuracy.

After the initial evaluation, a qualitative error analysis was conducted to identify common types of prediction errors. Based on this analysis, several modifications were made to the legal contract datasets to enhance the model's accuracy:

- **Lowercasing all input text:** This addressed issues with case mismatch between the training data and test data, which had previously led to incorrect predictions.
- **Simplifying questions:** Questions were reduced to their essential labels to avoid unnecessary complexity and ambiguity.
- **Dataset filtering:** Older and irrelevant contracts were removed to focus on modern legal terminology and improve the efficiency of the training process.

These modifications significantly improved model performance, particularly on the CUAD dataset, as discussed later in the Results section.

4. Data Collection

Several widely recognized datasets for both general-purpose QA tasks and domain-specific legal document analysis were utilized. The goal was to compare the performance of the ELECTRA model on both types of datasets, with a focus on legal contract review.

4.1 Datasets

SQuAD Dataset: SQuAD is a large-scale dataset consisting of over 100,000 questions based on Wikipedia articles (Rajpurkar et al., 2016). It serves as a benchmark for testing the reading comprehension abilities of QA models. The dataset provides exact answers from the text and is often used as a baseline for testing new models.

Adversarial QA Dataset: This dataset uses the same articles from SQuAD but incorporates questions designed to mislead the model (Jia & Liang, 2017). These questions introduce ambiguities or subtle changes in phrasing, making it more difficult for models to generate accurate answers. Testing on this dataset allowed us to evaluate how well ELECTRA performs when faced with adversarial input.

CUAD Dataset: CUAD consists of 510 legal contracts, labeled with over 13,000 annotations across 41 key legal clauses (Hendrycks et al., 2021). This dataset is particularly challenging due to the length and complexity of legal contracts, making it a valuable resource for testing models in the legal domain. CUAD is manually annotated and focuses on critical clauses that are relevant for legal contract review.

Filtered CUAD Dataset: A modified version of CUAD, the Filtered CUAD dataset excludes contracts signed before 2002 and reduces the number of contract categories from 41 to 12 (Apostolo, 2022). This refined dataset focuses on recent contract formats, which helps reduce confusion and improve the speed and accuracy of model predictions. It contains 385 legal contracts and over 13,000 labels.

CUAD QA Dataset: This variant of the CUAD dataset simplifies the questions to align directly with the clause labels. By trimming the questions to focus solely on the label form, this dataset helps reduce the cognitive load on the model, allowing for more efficient processing of legal clauses (Hao, 2022). It contains 11,178 training examples and 1,244 test examples.

5. Results and Discussion

The hyperparameters of the ELECTRA model were adjusted to achieve optimal performance across different datasets. The fine-tuning process focused on key parameters such as batch size, learning rate, and stride length during tokenization. These parameters were systematically altered and evaluated to determine their impact on performance. Table 1 shows the final set of hyperparameters that yielded the best results.

One of the most significant improvements was observed when the stride was adjusted during the tokenization process. Reducing the stride from 128 to 32 resulted in a 35% increase in the F1 score. This demonstrates how sensitive the model's performance is to changes in the length of the input tokens. However, this improvement was not uniform across all datasets. It was noted that smaller improvements were seen when working with legal documents, likely due to the complexity and length of these texts compared to simpler datasets like SQuAD.

The observed variation in performance improvement across datasets following stride adjustments was further analyzed. In the CUAD dataset, the structural complexity and extended clauses in legal documents posed challenges for the reduced stride, which excels in simpler datasets by improving token alignment granularity. Despite the potential for better token mapping, the dense syntax and embedded dependencies in legal texts limited the benefit of the stride change. Future work could explore adaptive stride mechanisms that dynamically adjust based on input text characteristics, potentially mitigating performance variability across diverse datasets.

Despite achieving strong results on some datasets, the model exhibited vulnerability to adversarial examples. This observation motivated further investigations with adversarial datasets and the introduction of custom adversarial triggers.

Table 1. Choices of hyperparameters for optimum results

Hyperparameter	Value
No. of Epochs	3
Batch size	8
Learning rate	2×10^{-5}
Stride	32
Max answer length	64

5.1 Performance on SQuAD and Adversarial QA Datasets

The initial evaluation was conducted on the SQuAD dataset, which is widely used to benchmark QA models. The ELECTRA model was initially trained on 97,714 examples of SQuAD and then evaluated on 10,626 examples of the same data set. The model performed well, achieving an Exact Match (EM) score of 78.6% and an F1 score of 86.5%. These values serve as the baseline for evaluating the model's overall performance on standard reading comprehension tasks.

Following this, the model was tested on the Adversarial QA dataset, which introduces more challenging and misleading questions. The model's performance dropped significantly when evaluated on the 3,000 examples of the Adversarial QA dataset. The EM score fell to 17.5%, and the F1 score decreased to 27.4%. This sharp decline in accuracy highlights the challenges posed by adversarial examples and demonstrates the limitations of the ELECTRA model when faced with ambiguity or misleading context. The drop in performance is due to the adversarial nature of the dataset, where questions are specifically designed to confuse the model. Many of the incorrect predictions were

related to syntactic confusion or wrong answer spans. The adversarial dataset showed that the model struggled with identifying the correct answer boundaries when questions were phrased in a misleading way. The results for both datasets are summarized in Table 2 (first two rows). These initial results emphasize the importance of exploring adversarial techniques to improve model resilience.

A universal adversarial trigger was then proposed to further examine the model's vulnerability. Unlike the more complex gradient-guided search methods proposed by (Wallace et al., 2019), the proposed trigger in this study relied on selecting words directly from the question to create confusion. This trigger was designed using a simple approach, wherein specific words from the question were added to the context to confuse the model. More specifically, the first two words and the last two words of the question, plus a fixed sentence, were added to the beginning or end of a context. The question's first two and last two words were selected as they often include a *wh*-word and some important information about the question asked. The question words were followed by a fixed sentence that was designed to include a person's name, activity, adverb, location, date, reason, and organization. The expectation was that combining the question words and the fixed sentences would confuse the model about the word connections and dependencies to pick the wrong words for the answer. One of the examples of such a fixed sentence added to the dataset was: "*because Lionel Messi played soccer aggressively in Hawaii on May 55, 2032 to support Vegan Organization*".

The proposed trigger was tested on the SQuAD dataset, and its effect was immediately noticeable. The EM score dropped from 78.6% to 70.6%, and the F1 score decreased from 86.5% to 77.2%. This demonstrated that even a simple adversarial modification could significantly degrade the model's performance. Table 2 (last row) presents the results with the adversarial trigger. The model's ability to accurately predict the correct answer decreased by approximately 10% in both the EM and F1 metrics. The trigger exploited the model's reliance on specific tokens for context understanding, causing confusion in the token selection process.

To better understand the model's performance under adversarial conditions, a quantitative analysis of error categories was performed. The results showed that the model failed to provide any prediction in 55% of cases and produced incorrect predictions in 35%. The remaining cases include ambiguous predictions, prediction formatting issues, and unclassifiable errors or edge cases. Among the incorrect predictions, 70% were due to mismatched spans, primarily caused by syntactic ambiguities in the adversarially perturbed contexts. This indicates that the adversarial triggers effectively exploit the model's dependency on specific token cues, disrupting the span selection process. Despite the addition of nonsensical fixed sentences in the triggers, the model successfully filtered these out, as none of the incorrect predictions included terms from these sentences. This highlights the model's capability to ignore irrelevant input while still being vulnerable to ambiguities introduced by adversarial perturbations.

Table 2. Test metrics on the SQuAD dataset

Approach	**EM (%)**	**Δ EM***	**F1 (%)**	**Δ F1***
Default Model (SQuAD)	78.6	-	86.5	-
Adversarial QA (Jia & Liang, 2017)	17.5	–78%	27.4	–71%
Proposed Adversarial Trigger	70.6	–10%	77.2	–11%

* % Change with respect to the default model

5.2 Performance on Legal Contract Datasets (CUAD and Variations)

The next set of experiments evaluated the model's performance on CUAD and its variations including filtered CUAD and CUAD QA. The CUAD dataset represents a significant departure from the simpler SQuAD dataset, as legal documents are often lengthy, complex, and full of jargon that complicates question answering tasks.

Initial training on the SQuAD dataset followed by testing on CUAD resulted on very poor performance markers, 1% EM and 6.2 % F1 scores. These poor results are once again attributed to the nature of the CUAD dataset containing very long contexts of real-world contracts that are not polished or modified according to everyday texts. The questions are often a statement followed by a synopsis of the form: "Highlight the parts (if any) of this contract related to 'Document Name' that should be reviewed by a lawyer. Details: The name of the contract". Further investigation also

revealed that even some of the benchmark answers provided for this dataset are wrong. These make it very difficult for the model to correctly understand the question and locate the answer in the context.

When trained on the same dataset, the ELECTRA model's performance on CUAD was increased yet substantially lower than on SQuAD. The model achieved an EM score of 30.8% and an F1 score of 44.2%. This indicates a significant reduction in accuracy when handling legal text, where identifying the correct answer spans is much more challenging due to the dense and ambiguous nature of the contracts. Many of the answers in CUAD are embedded within long clauses, and the questions are often vague or overly detailed, adding to the difficulty.

A breakdown of the types of errors for a sample of 100 wrong predictions on the CUAD dataset is shown in Figure 1. The errors are classified into the following categories. **Imprecise answer boundaries:** related to the selection of the wrong span for the answer by the model. **Syntactic complications & ambiguity:** related to the wrong interpretation of the context. **Paraphrase problems:** related to questions paraphrase a different part of the context. **Same entity type confusion:** related to entity confusion resulting in choosing the wrong meaning. **No prediction:** empty prediction is returned. **Correct prediction, wrong answer:** the correct prediction is considered wrong because the provided answers were wrong. **Casing mismatch:** related to a mismatch of letter casing between the answer and prediction.

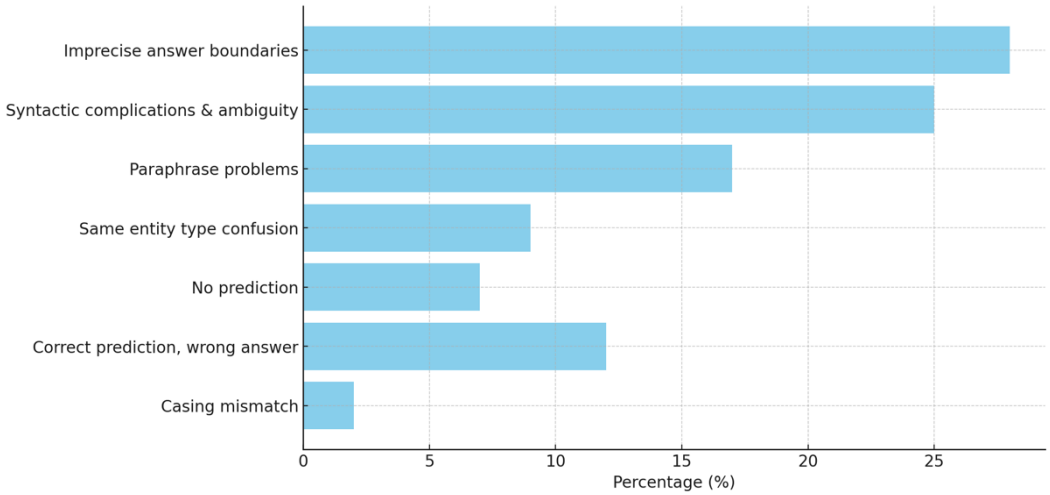
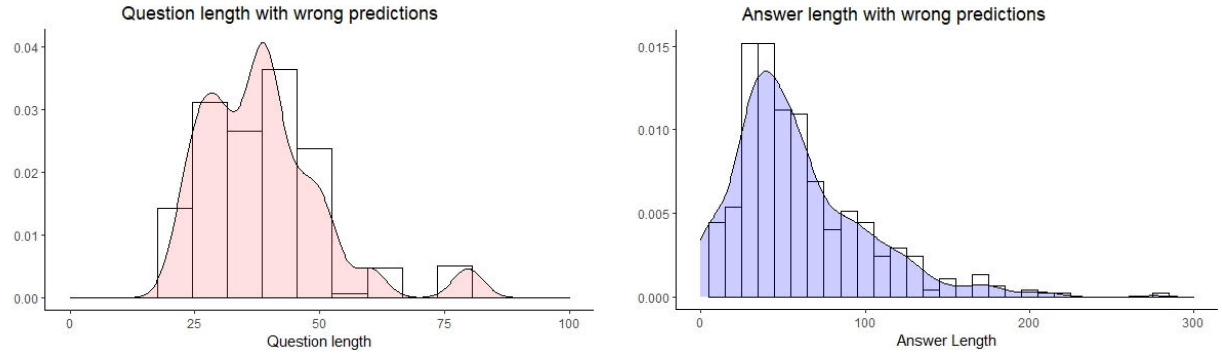


Figure 1. Breakdown of error types for a sample of 100 wrong samples of the CUAD dataset

Figure 2 shows the distribution of question lengths and answer lengths for the incorrectly predicted samples. These figures demonstrate that the errors were not necessarily tied to very long or very short questions. In fact, most incorrect predictions occurred for questions and answers of moderate length (around 38 words). This suggests that the complexity of the language, rather than the length, is a key factor in the model's performance on CUAD.



(a) (b)

Figure 2. Distribution of question lengths (a) and answer lengths for the incorrectly predicted samples.

5.3 Improving Model Performance

Given the poor performance on CUAD, several experiments were conducted to improve the model's accuracy. Three different preprocessing strategies were applied to the dataset:

- **Case 1:** The questions were shortened by removing the more complex, introductory parts and focusing solely on the key question.
- **Case 2:** The questions and contexts were lowercased before being fed into the model.
- **Case 3:** The entire input (questions, contexts, and answers) was lowercased.

Table 3 summarizes the results of these modifications. The first modification, which shortened the questions, had no significant impact on performance. The model was able to handle the original questions adequately, meaning that simplifying the question form did not reduce errors. However, lowercasing the questions and contexts (Case 2) showed a slight improvement in performance. This suggests that case mismatches between the training and test datasets contributed to the model's errors. The most substantial improvement came from lowercasing the entire input (Case 3), where the EM score improved by 20% and the F1 score by 12.2%. This demonstrates the importance of consistent case formatting in improving model accuracy.

Table 3. Test metrics of the adversarial methods on the SQuAD dataset

Approach	**EM (%)**	**Δ EM***	**F1 (%)**	**Δ F1***
Default Model (SQuAD)	30.8	-	44.2	-
Case 1 Modification	30.8	0%	44.2	0%
Case 2 Modification	31.2	+1.3%	44.8	–1.4%
Case 3 Modification	37.0	+20.1%	49.6	+12.2%

* % Change with respect to the default model

A breakdown of the types of errors for 100 samples of the wrong predictions on the modified CUAD dataset is shown in Figure 3. As shown, the share of case mismatch category of errors is now re-distributed among other categories.

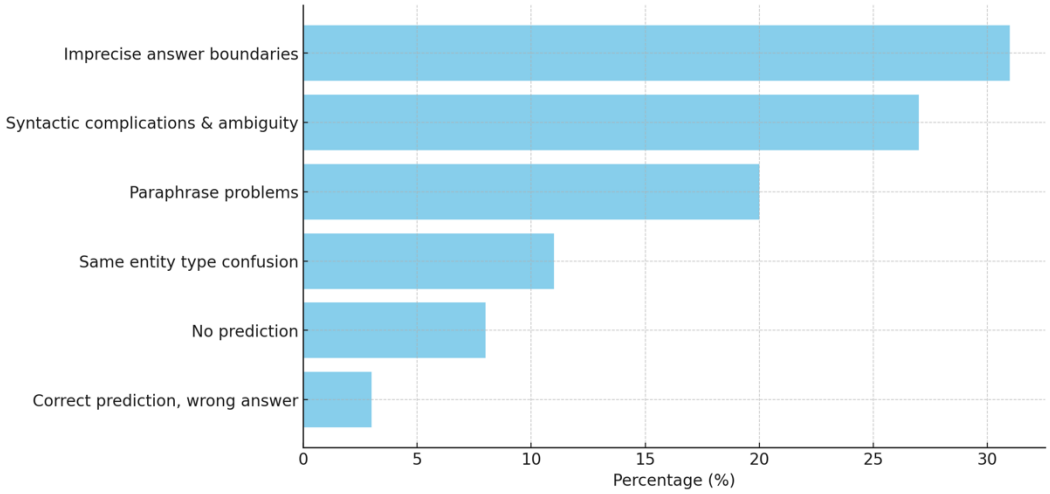


Figure 3. Breakdown of error types for wrong 100 samples of the modified CUAD dataset

5.4 Comparative Performance Across CUAD Variants

Using Case 3 modification, the model was then tested on Filtered CUAD and CUAD QA, two variants of the CUAD dataset. The Filtered CUAD dataset excludes older contracts and focuses on more recent legal agreements, while CUAD QA simplifies the questions to match the labels directly. As shown in Table 4, the model's performance on modified versions of these datasets improved on the original variants, though the gains were not as dramatic as those seen on the original CUAD dataset. The Filtered CUAD dataset yielded an EM score of 51.7% and an F1 score of 62.7%, an improvement over the original CUAD dataset. This is likely due to the removal of outdated contract formats, which simplified the task for the model.

Table 4. Test metrics across different CUAD variants

Dataset	**EM (%)**	**Δ EM***	**F1 (%)**	**Δ F1***
CUAD	30.8	-	44.2	-
Modified CUAD	37.0	+20.1%	49.6	+12.2%
Filtered CUAD	51.7	-	62.7	-
Modified Filtered CUAD	53.0	+2.5%	63.2	0.8%
CUAD QA	18.3	-	28.1	-
Modified CUAD QA	20.0	+9.3%	29.6	+5.3%

* % Changes of the modified variants (Case 3) compared to the original variants

The CUAD QA dataset, which simplifies the questions by reducing them to the form of labels, resulted in the lowest initial performance. The EM score was 18.3% and the F1 score was 28.1%. However, after applying the lowercasing modifications, the scores improved slightly to 20.0% for EM and 29.6% for F1. This suggests that the simplified question format in CUAD QA does not significantly help the model without additional improvements, such as better case matching and tokenization strategies.

6. Conclusion and Future Work

In this study, the performance of the ELECTRA model was evaluated on several Question-Answering (QA) datasets, including both general-purpose datasets like SQuAD and domain-specific legal contract datasets like CUAD. The results highlighted the model's ability to handle standard QA tasks effectively but also revealed significant challenges when it was applied to legal documents. The performance on the CUAD dataset demonstrated the complexities involved in analyzing real-world legal contracts, where the model achieved an Exact Match (EM) score of 30.8% and an F1 score of 44.2%. A universal adversarial trigger was introduced to further stress-test the model, which resulted in a performance degradation of approximately 10% on the SQuAD dataset. This underscores the model's vulnerability to adversarial inputs, which exploit weaknesses in its token selection process. Several improvements were explored to enhance the model's accuracy. Notably, lowercasing all input text led to significant improvements,

with a 20% increase in the EM score and a 12.2% increase in the F1 score on the CUAD dataset. This suggests that case mismatches between the training and testing data were a substantial source of error. Despite these gains, the model continued to struggle with complex legal language, especially in cases involving ambiguous phrasing or long, detailed clauses. The analysis of the model's performance on various CUAD variants, including Filtered CUAD and CUAD QA, revealed that simplifying the input data and standardizing the case could partially mitigate these issues. However, even with these modifications, the model's accuracy remains far from perfect, particularly in the legal domain, where the stakes for accuracy are high. Future work should focus on adaptive tokenization approaches, domain-specific fine-tuning, embedding-level adversarial defenses, and improving adversarial robustness. Additionally, expanding the dataset and enhancing model interpretability could help improve accuracy in legal contexts, making the model more reliable for high-stakes applications such as contract review.

Acknowledgments

This study is partially funded by the Kuwait Foundation for the Advancement of Sciences (KFAS) under grant no. CN2418TT2286.

References

Apostolo, A. *Filtered CUAD Dataset*. Https://Huggingface.Co/Datasets/Alex-Apostolo/Filtered-Cuad. 2022

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., & Roli, F. Evasion Attacks against Machine Learning at Test Time. In H. Blockeel, K. Kersting, S. Nijssen, & F. Železný (Eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 387–402, 2013.

Calijorne Soares, M. A., & Parreiras, F. S. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, *32*(6), 635–646, 2020. https://doi.org/10.1016/j.jksuci.2018.08.005

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *International Conference on Learning Representations*. 2019. https://openreview.net/forum?id=r1xMH1BtvB

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*. 2019 https://api.semanticscholar.org/CorpusID:52967399

Ganguly, D., Conrad, J. G., Ghosh, K., Ghosh, S., Goyal, P., Bhattacharya, P., Nigam, S. K., & Paul, S. *Legal IR and NLP: The History, Challenges, and State-of-the-Art* (pp. 331–340). 2023 https://doi.org/10.1007/978-3-031-28241-6_34

Goyal, S., Doddapaneni, S., Khapra, M. M., & Ravindran, B. A Survey of Adversarial Defenses and Robustness in NLP. *ACM Computing Surveys*, *55*(14s), 1–39. 2023 https://doi.org/10.1145/3593042

Hao, C. *CUAD QA*. Https://Huggingface.Co/Datasets/Chenghao/Cuad_qa. 2022

Hassan, F. ul, Le, T., & Lv, X. Addressing Legal and Contractual Matters in Construction Using Natural Language Processing: A Critical Review. *Journal of Construction Engineering and Management*, *147*(9). 2021 https://doi.org/10.1061/(ASCE)CO.1943-7862.0002122

Hendrycks, D., Burns, C., Chen, A., & Ball, S. *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*. 2021 arXiv preprint arXiv:2103.06268

Jia, R., & Liang, P. Adversarial examples for evaluating reading comprehension systems. *ArXiv Preprint ArXiv:1707.07328*. 2017

Kaliyar, R. K. A Multi-layer Bidirectional Transformer Encoder for Pre-trained Word Embedding: A Survey of BERT. *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 336–340. 2020 https://doi.org/10.1109/Confluence47617.2020.9058044

Kuchipudi, B., Nannapaneni, R. T., & Liao, Q. Adversarial machine learning for spam filters. *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 1–6. 2020 https://doi.org/10.1145/3407023.3407079

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., & Frossard, P. Universal Adversarial Perturbations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 86–94. 2017 https://doi.org/10.1109/CVPR.2017.17

Nassiri, K., & Akhloufi, M. Transformer models used for text-based question answering systems. *Applied Intelligence*, *53*(9), 10602–10635. 2023 https://doi.org/10.1007/s10489-022-04052-8

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. 2016 arXiv preprint arXiv:1606.05250

Shao, K., Zhang, Y., Yang, J., Li, X., & Liu, H. The triggers that open the NLP model backdoors are hidden in the adversarial samples. *Computers & Security*, *118*, 102730. 2022 https://doi.org/10.1016/j.cose.2022.102730

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. *Universal Adversarial Triggers for Attacking and Analyzing NLP*. 2019 arXiv preprint arXiv:1908.07125

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. Adversarial Attacks on Deep-learning Models in Natural Language Processing. *ACM Transactions on Intelligent Systems and Technology*, *11*(3), 1–41. 2020 https://doi.org/10.1145/3374217

Biographies

Amir Barakati is an Associate Professor in the Department of Mechanical and Aerospace Engineering at George Washington University. With a diverse academic background, he earned his PhD in Mechanical Engineering from the University of Iowa, an MSc in Data Science from the University of Texas at Austin, an MSc in Aerospace Engineering from Sharif University of Technology, and a BSc in Aerospace Engineering from Amirkabir University of Technology. His research interests encompass machine learning, engineering education, biocomposites, computational mechanics, and smart material systems, with a particular focus on sustainable engineering solutions and innovative educational practices.

Arash Mehdizadeh is an Assistant Professor in the Electrical and Electronics Engineering Department at the Australian University in Kuwait. He holds a PhD in Electrical and Electronics Engineering from The University of Adelaide, Australia, where he earned a Dean's Commendation for his doctoral thesis. He then completed his Post-Doctoral Fellowship at the University of Western Australia. He also holds an MSc in Data Science from The University of Texas at Austin and an MSc and a BSc in Computer Systems Engineering from Amirkabir University of Technology. He specializes in data science, artificial intelligence, microelectronics, and biomedical engineering, with a strong focus on computational biology and embedded systems design. His research interests span various domains, including machine learning, computational biology, microelectronics, and biomedical devices. He has contributed extensively to the field of energy-efficient building solutions, biomedical implants, and advanced computational modeling. Throughout his academic career, he has secured multiple research grants, including from the Kuwait Foundation for the Advancement of Sciences (KFAS), and has published in prestigious journals and conferences. His work on artificial intelligence applications and computational systems has earned him several best project awards at the Australian University in Kuwait. In addition to his research contributions, Dr. Mehdizadeh is actively involved in teaching and mentoring students, focusing on project-based learning. He is a professional member of IEEE, Engineers Australia, and The Chartered Institute for IT.

Baktanoosh Nakhjavani is currently a Project Engineer at FATA Automation, a leading automation company specializing in the design and fabrication of robotics and automated production lines across various industries, including automotive. He holds a bachelor's and master's degree in aerospace engineering from IUST, a bachelor's degree in computer science from the University of Maryland, and a master's degree in data science from The University of Texas at Austin. Baktanoosh combines a robust technical foundation with a strong analytical mindset, honed through years of experience in engineering roles. He has a keen interest in research areas such as reinforcement learning, machine learning, the application of machine learning and image processing in healthcare and dentistry, and the use of AI in robotics and automation.