Neural Networks 194 (2026) 108105

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Deceiving question-answering models: A hybrid word-level adversarial approach

Jiyao Li [a], Mingze Ni [a], Yongshun Gong [b], Wei Liu [a,*]

[a] University of Technology Sydney, 15 Broadway, Sydney, 2007, NSW, Australia
[b] Shandong University, 27 Shanda Nanlu, Jinan, 250100, Shandong, China

ARTICLE INFO

Keywords:
Neural language processing
Adversarial attack
Deep learning
Question answering

ABSTRACT

Deep learning underpins most of the currently advanced natural language processing (NLP) tasks such as textual classification, neural machine translation (NMT), abstractive summarization and question-answering (QA). However, the robustness of the models, particularly QA models, against adversarial attacks is a critical concern that remains insufficiently explored. This paper introduces QA-Attack (Question Answering Attack), a novel word-level adversarial strategy that fools QA models. Our attention-based attack exploits the customized attention mechanism and deletion ranking strategy to identify and target specific words within contextual passages. It creates deceptive inputs by carefully choosing and substituting synonyms, preserving grammatical integrity while misleading the model to produce incorrect responses. Our approach demonstrates versatility across various question types, particularly when dealing with extensive long textual inputs. Extensive experiments on multiple benchmark datasets demonstrate that QA-Attack successfully deceives baseline QA models and surpasses existing adversarial techniques regarding success rate, semantics changes, BLEU score, fluency and grammar error rate.

1. Introduction

Question-answering (QA) models, a key task within Sequence-to-Sequence (Seq2Seq) frameworks, aim to enhance computers' ability to process and respond to natural language queries. As these models have evolved, they have been widely adopted in real-world applications such as customer service chatbots (Nuruzzaman & Hussain, 2018), search engines (Zhu et al., 2021), and information retrieval in fields like medicine (Jin et al., 2021) and law (Martinez-Gil, 2023). However, despite the significant progress in deep learning and natural language processing (NLP), these models remain vulnerable to adversarial examples, leading to misinformation, privacy breaches, and flawed decision-making in critical areas (Dong et al., 2022; Hathaliya et al., 2022; Klopfenstein et al., 2017; Sun et al., 2021; Yin et al., 2018). This highlights the importance of understanding how adversarial examples are generated from the attackers' perspective and potential defense mechanisms – an area that remains under-explored.

QA models are expected to comprehend given texts and questions, providing accurate and contextually relevant answers (Soares & Parreiras, 2020). These models primarily address two types of questions: Informative Queries and Boolean Queries. The Informative Queries typically begin with interrogative words such as "who," "what," "where," "when," "why," or "how," requiring detailed and specific information from the provided context. Although models like T5 (Raffel et al., 2020), LongT5 (Guo et al., 2022), and BART (Lewis et al., 2020), which follow an encoder-decoder structure, have demonstrated strong performance, they still suffer from maliciously crafted adversarial examples. Initially, studies like "Trick Me If You Can" (Wallace et al., 2019b) primarily relied on human annotators to construct effective adversarial question-answering examples. This methodology, however, inherently constrains scalability and increased resource demands. As research progressed, automated approaches for attacking textual classifiers in QA models emerged. Gradient-based methods, as employed in Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015), RobustQA (Yasunaga et al., 2018), UAT (Wallace et al., 2019a), and HotFlip (Ebrahimi et al., 2018), were developed to identify and modify the most influential words affecting model answers. Building upon a deeper understanding of QA tasks, subsequent studies explored more targeted strategies. For instance, Position Bias (Ko et al., 2020), TASA (Cao et al., 2022), and Entropy Maximization (Shinoda et al., 2023) investigated the manipulation of sentence locations and the analysis of answer sentences to identify vulnerable parts of the context. These approaches refined the attack methods by applying modifications through paraphrasing or replacing original sentences, thus enhancing the effectiveness of adversarial examples. However, these methods encounter two primary challenges: 1) None of these attack methods is suitable for both "informative queries" and

* Corresponding author.
E-mail addresses: jiyao.li-1@student.uts.edu.au (J. Li), mingze.ni@uts.edu.au (M. Ni), ysgong@sdu.edu.cn (Y. Gong), wei.liu@uts.edu.au (W. Liu).

https://doi.org/10.1016/j.neunet.2025.108105
Received 17 November 2024; Received in revised form 21 July 2025; Accepted 8 September 2025
Available online 13 September 2025

"boolean queries". 2) Constraining the search space for optimal vulnerable words to answer-related sentences compromises attack effectiveness; meanwhile, targeting entire sentences proves inefficient (Jia & Liang, 2017).

In addition, Boolean Queries seek a simple binary "Yes" or "No" answer. Models like BERT (Devlin et al., 2019), RoBERTa (Zhuang et al., 2021), and GPT variants (Antaki et al., 2024; Bongini et al., 2022; Klein & Nabi, 2019; Stiennon et al., 2020), which excel at sentence-level understanding and token classification, are widely used for Boolean QA tasks. These models leverage their deep contextual understanding of language to accurately determine whether a given statement is true or false, making them state-of-the-art baselines for the task. Researchers have proposed various approaches to target boolean classifiers in the context of Boolean Queries attacks. Attacks like (Garg & Ramakrishnan, 2020; Jin et al., 2020; Li et al., 2020; Ren et al., 2019; Zang et al., 2020), which involve adding, relocating, or replacing words, are based on the influence that each word has on the prediction. They retrieve word importance by the output confidence to the level or with gradient. However, gradient calculation is computationally intensive and ineffective when dealing with long context input, and knowing victim models' internal information is unrealistic in practice.

We present QA-Attack, an adversarial attack framework tailored for both Informative Queries and Boolean Queries in QA models. It is especially suitable for advancing the research in semi-black-box and gray-box attacks. QA-Attack uses a Hybrid Ranking Fusion (HRF) algorithm that integrates two methods: Attention-based Ranking (ABR) and Removal-based Ranking (RBR). ABR identifies important words by analyzing the attention weights during question processing, while RBR evaluates word significance by observing changes in the model's output when specific words are removed. The HRF algorithm combines these insights to locate vulnerable tokens, which are replaced with carefully selected synonyms to generate adversarial examples. These examples mislead the QA system while preserving the input's meaning. This unified attack method improves both performance and stealth, ensuring realistic applicability for both types of queries. In summary, our work makes the following key contributions:

- We present QA-Attack with a Hybrid Ranking Fusion (HRF) algorithm designed to target question-answering models. This novel approach integrates attention and removal ranking techniques, accurately locating vulnerable words and fooling the QA model with a high success rate.
- Our QA-Attack can effectively target multiple types of questions. This adaptability allows our method to exploit vulnerabilities across diverse question formats, which significantly broadens the scope of potential attacks in various real-world scenarios.
- QA-Attack generates adversarial examples by implementing subtle word-level changes that preserve both linguistic and semantic integrity while minimizing the extent of alterations, and we conduct extensive experiments on multiple datasets and victim models to thoroughly evaluate our method's effectiveness in attacking QA models.

The rest of this paper is structured as follows. We first review QA system baselines and adversarial attacks for QA models in Section 2. Then, we detail our proposed method in Section 3. We evaluate the performance of the proposed method through extensive empirical analysis in Section 4. We conclude the paper with suggestions for future work in Section 5.

2. Related work

This section provides a comprehensive overview of question-answering models and examines the existing research on adversarial attacks against them.

2.1. *Question answering models*

Question answering represents a complex interplay of NLP, information retrieval, and reasoning capabilities (Soares & Parreiras, 2020; Yigit & Amasyali, 2024). Basically, these models are designed to process an input question and a context passage, extracting or generating an appropriate answer through elaborate analysis of the semantic relationships between these elements (Wang, 2022). Modern QA systems typically rely on deep learning models with transformer-based architectures like BERT (Devlin et al., 2019) and its variants (Lan et al., 2020; Sanh et al., 2019; Zhuang et al., 2021) being particularly prevalent. These models excel at capturing contextual information and understanding nuanced relationships in the text with transformers, allowing them to perform impressively on QA tasks. In addition to these transformer models, encoder-decoder architectures such as T5 (Khashabi et al., 2020; Raffel et al., 2020) and BART (Lewis et al., 2020), GPT (Brown et al., 2020) and PEGASUS (Zhang et al., 2020) have also become prominent in QA models. These models utilize an encoder to process the input question and context, transforming them into a rich, context-aware representation, and the decoder is then used to generate a coherent and contextually appropriate answer.

2.2. *Previous works on attacking QA models*

With the development of NLP techniques, recent research has increasingly focused on developing sophisticated textual adversarial examples for QA systems (Wallace et al., 2019b). The inherent differences between "informative queries" and "boolean queries" necessitate distinct attacking diversities due to their unique answer structures (Wallace et al., 2019a). Attacks on boolean QA pairs closely resemble methods used to mislead textual classifiers. These attacks primarily operate at the word level, aiming to manipulate the model's binary (yes/no) output (Garg & Ramakrishnan, 2020; Li et al., 2020). In contrast, informative queries present a more complex challenge. These attacks frequently target the sentence level, requiring an approach to disrupt the model's comprehensive understanding (Li et al., 2021).

2.2.1. *Boolean queries attacks*

Boolean queries are similar to classification tasks in NLP, while the answer is based on two-way input: question and context. They are vulnerable to attacks designed for NLP classifiers when question and context are simply encoded and concatenated. Approaches such as Garg and Ramakrishnan (2020), Jin et al. (2020), Li et al. (2020), Madry et al. (2018), Ren et al. (2019), Zang et al. (2020) concentrate on altering individual words based on their influence on model predictions. These methods typically employ carefully selected synonyms for word substitution. The process of word replacement is guided either by the direct use of BERT Masked Language Model (MLM) (Devlin et al., 2019) or by leveraging gradient information to determine optimal substitution candidates. While effectively fool classifiers (boolean queries), these attacks were initially designed for classification tasks and have shown limited efficacy when applied to the question-and-context format of QA systems. To address this limitation, some attack methods for Seq2Seq models have been adapted for QA models. UAT (Wallace et al., 2019a), which averages gradients and modifies input data to maximize the model's loss, has been adapted for QA but still struggles with boolean queries due to their simplicity. Similarly, TextBugger (Li et al., 2019), which focuses on character-level perturbations, also faces challenges in handling the deeper semantic understanding required in QA, especially for multi-sentence reasoning. Liang's approach (Liang et al., 2018), relying on confidence-based manipulations, has difficulty reducing the model's certainty in boolean queries where the binary answers leave less room for variation in confidence. Although these approaches offer improved accuracy in attacking informative questions with minor modifications, they struggle with boolean queries. We argue that these methods face challenges in identifying the most vulnerable words when dealing with concatenated question-context input relationships. The MLQA attack (Rosenthal et al., 2021) attempts to bridge this gap by utilizing attention weights to identify and alter influential words. However, this method, developed specifically for multi-language BERT models, may not fully address QA-specific vulnerabilities.

2.2.2. Informative queries attacks

In contrast to boolean queries, adversarial attacks on informative queries within QA systems share fundamental similarities with attacks on other Seq2Seq models (Bahdanau et al., 2014; Cheng et al., 2020; Li & Liu, 2023; Luong et al., 2015; Moosavi-Dezfooli et al., 2016; Ribeiro et al., 2018), concentrating more on the inter-relationship between question and context. The defense mechanisms like RobustQA (Yasunaga et al., 2018) have been developed to enhance model resilience through improved training methods, and sophisticated attacks continue to successfully compromise these systems, especially when employing subtle manipulations of key input elements. Character-level attack methods, notably HotFlip (Ebrahimi et al., 2018), have demonstrated significant success by strategically flipping critical characters based on gradient information, leading to misinterpreting informative inputs. In the multilingual domain, MLQA (Talmor & Berant, 2019) leverages attention weights to identify and target crucial words, though its attention mechanism, primarily designed for multilingual functionality, may not fully exploit the intricate vulnerabilities within the model's attention architecture. Advanced techniques have emerged to target the influence that answers have on QA systems. Position Bias and Entropy Maximization methods exploit model weaknesses by manipulating contextual patterns and answer positioning, particularly effective in scenarios involving complex, lengthy responses. Syntactically Controlled Paraphrase Networks (SCPNs) (Iyyer et al., 2018) generate adversarial examples through strategic syntactic alterations while preserving semantic meaning. TASA (Targeted Adversarial Sentence Analysis) (Cao et al., 2022) primarily relies on manipulating the answer sentences to mislead QA models, making it particularly effective for informative queries where complex responses provide more opportunities for subtle modifications. However, this approach is not suitable for boolean queries, as the simplicity of yes/no answers limits the sentence-level manipulations that TASA depends on.

Despite significant progress in adversarial attacks on QA systems, existing methods still face several notable limitations. Most approaches are specialized for either informative or boolean queries, but not both. For instance, sentence-level attacks such as TASA are designed for informative QA tasks that require detailed answers, and are not applicable to boolean queries due to their simplicity and limited answer structure. In contrast, Textfooler primarily targets classification tasks and often fails to generalize well to open-ended or extractive QA formats. Additionally, sentence-level adversarial strategies, while often effective in misleading QA models, typically suffer from high computational cost due to the need for candidate generation, semantic similarity evaluation, and sentence rewriting. More critically, these approaches often introduce significant semantic shifts, which can alter the original intent of the context or question. This undermines the imperceptibility of adversarial examples

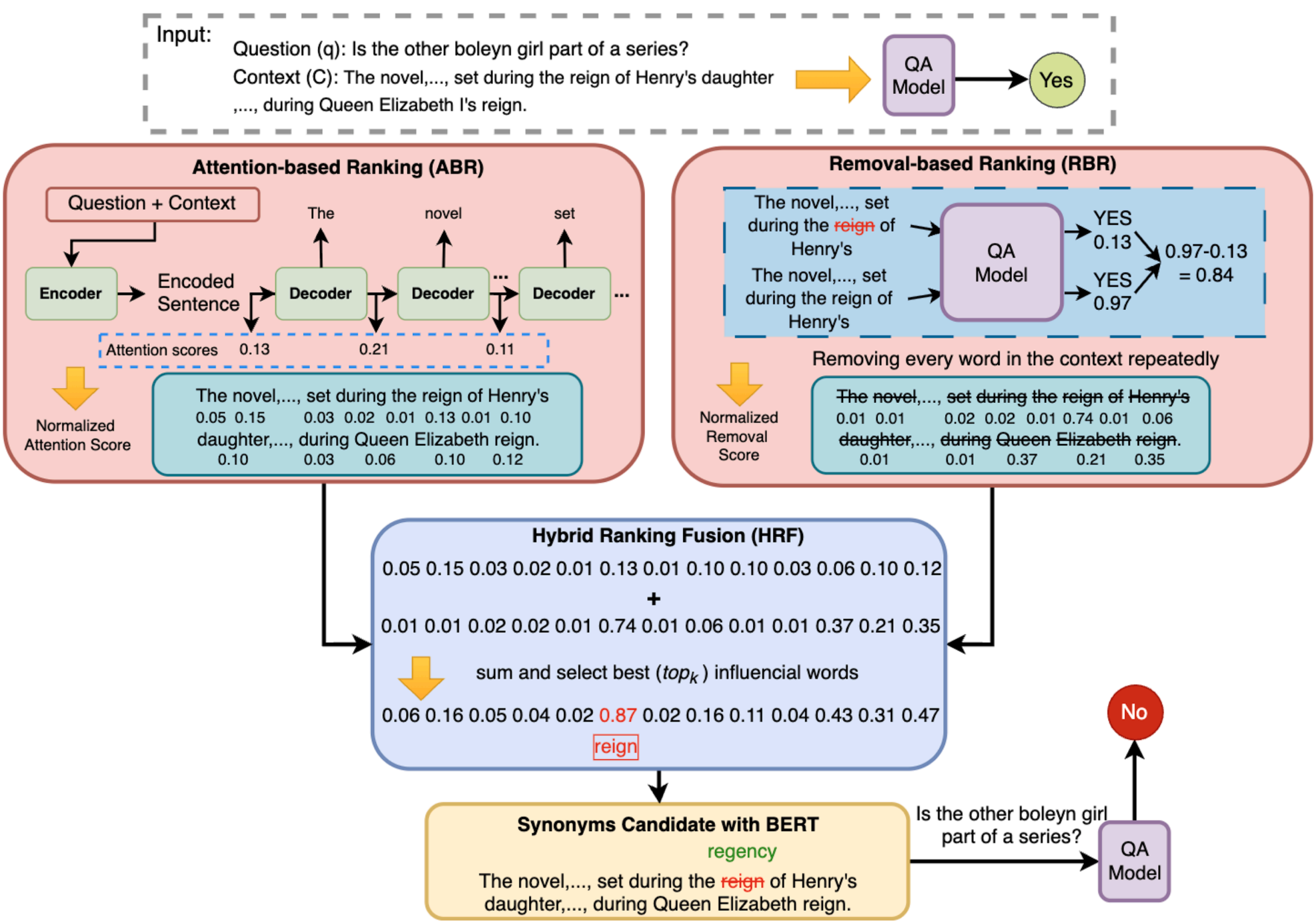


Fig. 1. The workflow of our QA-Attack algorithm for QA models. It processes question-context pairs through two parallel modules: Attention-based Ranking (ABR) and Removal-based Ranking (RBR). These modules generate attention and removal scores respectively for each word using customized attention mechanisms and removal ranking strategies. The scores are then aggregated, and the top_k highest-scoring words are selected as candidates. Finally, these candidates are replaced with BERT-generated synonyms to create adversarial examples that can effectively mislead the QA model.

and raises concerns about the validity and naturalness of the generated inputs, particularly in scenarios where preserving the original meaning is crucial.

Recent research has proposed several defense strategies to enhance the robustness of QA systems against adversarial attacks. Frequency-Guided Word Substitution (FGWS) (Mozes et al., 2021) improves input stability by substituting rare words with high-frequency synonyms during training, thereby strengthening the model's resilience to word-level perturbations. Similarly, Random Masking Training (RanMASK) (Zeng et al., 2023) enhances robustness by randomly masking tokens during fine-tuning, which encourages the model to generalize effectively over incomplete or partially corrupted inputs. These defense mechanisms make QA systems less vulnerable to various types of adversarial attacks.

3. Our proposed attack method

In this section, we introduce the QA-Attack algorithm. It can be summarized into three main steps. First, the method effectively captures important words in context by processing pairs of questions and corresponding context using attention-based and removal-based ranking approaches. Then, attention and removal scores are combined, allowing the identification of the most influential words. At last, a masked language model (Devlin et al., 2019) is utilized to identify potential synonyms that could replace the targeted words. The overall workflow of QA-Attack is shown in Fig. 1. In the following sections, we explain our model in detail.

3.1. Problem setting

Given a pre-trained question-answering model F, which receives an input of context C, question q, and outputs answer a, such that $F(q,C)=a$. The objective is to deceive the performance of F with perturbed context C' such that $F(q,C') \neq a$. To craft C', a certain number of perturbation c_{adv} is added to the context C by replacing some of its original tokens $\{c_1, c_2, \dots, c_n\}$.

3.2. Attention-based ranking (ABR)

Attention mechanisms were first used in image feature extraction in the computer vision field (Galassi et al., 2021; Lyu et al., 2023; Xu et al., 2015; Yang et al., 2020). However, they were later employed by Bahdanau et al. (2014), Ni et al. (2022) to solve machine translation problems. In translation tasks, attention mechanisms enable models to prioritize and focus on the most relevant parts of the input data (Luong et al., 2015). In question-answering tasks, attention scores are imported to examine the relationships between question and context, allowing the model to determine which words or phrases are most relevant to answering the question (Xiao & Zhu, 2023). Hence, we leverage the attention score to identify target words for our attack. We employ the attention mechanism from T5 (Raffel et al., 2020) that has been specifically optimized for question-answering tasks in UnifiedQA (Khashabi et al., 2020). As shown in Fig. 1, the "Attention-based Ranking" begins by encoding the input context and question through an encoder. During the encoding process, self-attention allows the model to analyze how each word in the input relates to every other word, effectively highlighting the words that carry the most weight in understanding both question and context. In the decoding process, cross-attention further refines this by focusing on the parts of the input most relevant to generating the correct output. By averaging the attention scores of all layers and heads, we match them to each input word.

The implement details are shown in Algorithm 1. The question & context pair is fitted into attention network A, and we filter out the attention scores for context (lines 1 to 8 of Algorithm 1). Then, the attention score of each word corresponding to each layer is summed up. After averaging and normalization, the word-level attention score is obtained.

Algorithm 1: QA-attack algorithm (adversarial generation).

Input : QA victim model $F(\cdot)$, logits L, question q, context C, words in the context c, reference answer a, attention network A, top k words to attack top_k, number of synonyms d, BERT MLM model $BERT$ for generating synonyms.
Output: Optimal adversarial sample C'

1 // **Attention-based Ranking**;
2 Compute attention scores: $\alpha \leftarrow [(c, A(q+C))]$;
3 Initialize attention score list: $attention\_scores \leftarrow [\ ]$;
4 **for** *each score in* α **do**
5   **if** $score \in C$ **then**
6   Append $score$ to $attention\_scores$;
7   **end**
8 **end**
9 // **Removal-based Ranking**;
10 Initialize importance score list: $importance\_scores \leftarrow [\ ]$;
11 **for** *each c in C* **do**
12   Generate modified context: $C^* \leftarrow C$ excluding c;
13   Compute importance score: $importance\_scores.append(|F(q,C^*) - F(q,C)|)$;
14 **end**
15 // **Hybrid Ranking Fusion**;
16 Combine attention and importance scores: $combined\_scores \leftarrow attention\_scores \cup importance\_scores$;
17 Select top_k words: $top\_k\_list \leftarrow \text{sort}(combined\_scores)[\ :\ top_k]$;
18 Initialize adversarial examples list: $Adv\_list \leftarrow [\ ]$;
19 **for** *each t in top_k_list* **do**
20   Generate adversarial token from d potential synonyms: $c_{adv} \leftarrow BERT(t)$;
21   Create adversarial context: $\Delta C \leftarrow [c_1, \dots, c_{adv}, \dots, c_n]$;
22   Append ΔC to Adv_list;
23 **end**

3.3. Removal-based ranking (RBR)

Previous studies on adversarial attacks in the text have shown that each word's significance can be quantified using an importance score (Cao et al., 2022; Jin et al., 2020; Li & Liu, 2023; Li et al., 2020). This score is largely determined by how directly the word influences the final answer. To enhance the efficacy of ranking progress, we rank each word in the context to obtain the removal importance score (lines 9 to 14 of Algorithm 1). Given the input context C containing n words from c_1 to c_n and question q, the importance score (removal score) of the i th $(1 \le i \le n)$ word c_i is:

$$I_i = L_F(a \mid q, C) - L_F(a \mid q, C \setminus c_i), \tag{1}$$

where $C \setminus c_i$ represents the context after deleting c_i, and $L_F = \log P(a \mid q, C)$ refers to the probability (logits) of the label, respectively.

3.4. Hybrid ranking fusion (HRF)

The attention-based and removal-based word selection techniques offer complementary perspectives on token significance, each highlighting different aspects of word importance. Consequently, we tend to choose words that both methods consider significant. This is achieved by adding the scores from each method for every word to create a fusion score.

When generating a fusion score, we address several key factors. First, we independently normalize the attention and removal scores before adding them together. Then, to balance attack effectiveness and efficiency, we introduce a top_k parameter, a positive integer that controls the number of words targeted. Finally, we select the top_k highest-scoring words for modification (lines 15 to 18 of Algorithm 1).

3.5. Synonym selection

Various synonym generation methods exist, including Word2Vec (Mikolov et al., 2013), Hownet (Dong & Dong, 2003), and WordNet (Miller, 1992). We adopt BERT (Devlin et al., 2019) for synonym selection due to its textual capabilities, which enable it to generate synonyms based on the complete sentence structure. Unlike Word2Vec's static embeddings or WordNet's fixed synonym lists, BERT's context-sensitive approach allows for dynamic synonym selection that preserves both semantic meaning and grammatical correctness. This contextual awareness makes BERT particularly effective for crafting natural and semantically coherent adversarial examples.

We process each selected word in the context by replacing it with the "[MASK]" token. This modified context is then input into the BERT Masked Language Model (MLM) to predict the most likely substitutions for the masked word. To expand the range of potential samples, we introduce a parameter d that controls the number of synonym substitutions considered (lines 19 to 23 of Algorithm 1). This approach allows us to generate a diverse set of imperceptible replacements while maintaining contextual relevance.

3.6. Candidate selection

We define an optimal adversary as one that maximizes the divergence between the predicted and attacked answers. For boolean queries (yes/no), we follow previous successful textual classifier approaches by comparing the logits of output labels. For informative queries, we aggregate the logits across individual words in the response. The optimal adversary C' is identified from the "Adv_list" using the logits derivation function L, as detailed in Algorithm 2.

Algorithm 2: QA-attack algorithm (optimization).

Initialize maximum gap: max_gap $\leftarrow -\infty$;
Initialize optimal adversarial context: $C' \leftarrow \emptyset$;
for *each* adv *in* Adv_list **do**
  if $F(adv) \neq a$ **then**
    Compute gap: gap $\leftarrow L(F(adv)) - L(F(C))$;
    if *gap* $>$ *max_gap* **then**
      Update maximum gap: max_gap $\leftarrow$ gap;
      Update optimal adversarial context: $C' \leftarrow adv$;
    end
  end
end
return Optimal adversarial sample C'

4. Experiment and analysis

In this section, we present a comprehensive evaluation of QA-Attack's performance compared to current state-of-the-art baselines. Our analysis covers several key aspects with various metrics, providing a thorough understanding of our method's capabilities, limitations, and performance across diverse scenarios. We provide a detailed analysis of attack performance and imperceptibility (Section 4.4). Besides, to gain deeper insights, we conduct ablation studies (Section 4.5) and assess attacking efficiency (Section 4.6). In addition, we examine QA-Attack's response to defense strategies (Section 4.8), exploring the effects of adversarial retraining (Section 4.7) and investigating the transferability of attacks (Section 4.9). Finally, we report the preference of our attack by investigating parts of speech preference (Section 4.10) and analyzing its robustness versus the scale of pre-trained models (Section 4.11).

4.1. Datasets and victim models

We assess QA-Attack using four *informative query* datasets: SQuAD 1.1 (Rajpurkar et al., 2016), SQuAD V2.0 (Rajpurkar et al., 2018), NarrativeQA (Kočiský et al., 2018), and NewsQA (Trischler et al., 2017), as well as two domain-specific datasets, EMRQA (Pampari et al., 2018) and FinQA (Chen et al., 2022), along with a *boolean query* dataset, BoolQ (Clark et al., 2019). Note that EMRQA and FinQA are domain-specific datasets designed to evaluate attack performance in specialized fields, and thus were excluded from our ablation experiments.

- SQuAD 1.1: Questions formulated by crowd workers based on Wikipedia articles. Answers are extracted as continuous text spans from the corresponding passages.
- SQuAD 2.0: Extension of SQuAD 1.1 incorporating unanswerable questions. These questions are designed such that no valid answer can be located within the provided passage, adding complexity to the task.
- NarrativeQA: Questions based on entire books or movie scripts. Answers are typically short and abstractive, demanding deeper comprehension and synthesis of narrative elements.
- NewsQA: Questions based on CNN news articles designed to test reading comprehension in the context of current events and journalistic writing.
- BoolQ: Dataset of boolean (yes/no) questions derived from anonymized, aggregated queries submitted to the Google search engine, reflecting real-world information-seeking behavior.
- EMRQA: Large-scale question-answer pairs generated from electronic medical records. It emphasizes domain-specific reasoning and medical information extraction, challenging models with clinical contexts.
- FinQA: Financial QA dataset focusing on numerical reasoning over semi-structured tables and text passages, requiring models to perform calculations and logical inference.

Our experiment includes three question-answering models for comparison. They are T5 (Khashabi et al., 2020), LongT5 (Guo et al., 2022), and $\text{BERT}_{\text{base}}$ (Devlin et al., 2019). The LongT5 is an extension of T5 with an encoder-decoder specifically for long contextual inputs. The BERT-based models are structured with bidirectional attention, meaning each word in the input sequence contributes to and receives context from both its left and right sides. Table 1 presents the distribution of dataset splits and F1 scores reported on each QA baseline.

4.2. Baseline attacks

For our experimental baselines, we employ seven leading attack methods: TASA (Cao et al., 2022), RobustQA (Yasunaga et al., 2018), Tick Me If You Can (TMYC)(Wallace et al., 2019b), T3(Wang et al., 2020), TextFooler (Jin et al., 2020), PIA (Parry et al., 2024), and LLM-Attack (Wang et al., 2023). We utilize the official implementation of T3 in its black-box setting, while TASA, TMYC, and RobustQA are employed with their standard configurations. TextFooler, originally not designed

Table 1
Dataset distribution and corresponding baseline performance (F1).

Dataset	Data distribution				Model performance (F1)		
	Total	Train	Validation	Test	T5	LongT5	$\text{BERT}_{\text{base}}$
SQuAD 1.1	100k	87,600	10,570	N/A	88.9	89.5	88.5
SQuAD V2.0	150k	130,319	11,873	N/A	81.3	83.2	74.8
NewsQA	119k	92,549	5165	5126	66.8	67.2	60.1
BoolQ	16k	9427	3270	3245	85.2	86.1	80.4
NarrativeQA	45k	32,747	3461	10,557	67.5	68.9	62.1
EMRQA	400k	300,000	50,000	50,000	71.2	72.5	69.3
FinQA	10k	7000	1500	1500	77.8	78.6	74.2

Table 2
Comparison of original and adversarial contexts for two types of queries. The table highlights the differences between the original and adversarial contexts, as well as the corresponding answers provided by the model before and after the attack.

Question	Was the movie "The Strangers" based on a true story?
Context	The Strangers is a 2008 American slasher film written and directed by Bryan Bertino. Kristen (Liv Tyler) and James (Scott Speedman) are expecting a relaxing weekend at a family vacation home, but their stay turns out to be anything but peaceful as three masked torturers leave Kristen and James struggling for survival. Writer-director Bertino was inspired by real-life events: the Manson family Tate murders, a multiple homicide; the Keddie Cabin Murders, that occurred in California in 1981; and a series of break-ins that occurred in his own neighborhood as a child.
Adversary	The Strangers is a 2008 American slasher thriller written and directed by Bryan Bertino. Kristen (Liv Tyler) and James (Scott Speedman) are spending a relaxing weekend at a family vacation home, but their stay turns out to be anything but peaceful as three masked torturers leave Kristen and James struggling for survival. Writer-director Bertino was influenced by real-life incidents: the Manson family Tate murders, a multiple homicide; the Keddie Cabin Murders, that occurred in California in 1981; and a series of break-ins that occurred in his own home as a child.
Original Answer	Yes
Attacked Answer	No
Question	Who ruled the Duchy of Normandy?
Context	The Normans were famed for their martial spirit …The Duchy of Normandy, which they formed by treaty with the French crown, was a great fief of medieval France, and under Richard I of Normandy was forged into a cohesive and formidable principality in feudal tenure …Norman adventurers founded the Kingdom of Sicily …an expedition on behalf of their duke, William the Conqueror, led to the Norman conquest of England at the Battle of Hastings in 1066.
Adversary	The Normans were famed for their warrior spirit …The Duchy of Normandy, which they formed by treaty with the French crown, was a great fief of medieval France, and under William I of Normandy was forged into a cohesive and formidable principality in feudal tenure …Norman adventurers invaded the Kingdom of Sicily …an invasion on behalf of their duke, William the Conqueror, led to the Norman conquest of England at the siege of Hastings in 1066.
Original Answer	The French crown
Attacked Answer	William I of Normandy

for question-answering tasks, was adapted for our experiments by modifying it to process the context only (questions are removed). For PIA and LLM-Attack, we employ their official implementations aligned with question answering tasks.

4.3. Experiment settings and evaluation metrics

The base setting of our experiments is to let $top_k = 5$, $d = 2$, and use a BERT-base-uncased[1] with 12 Transformer encoder layers (L) and 768 hidden layers (H) as the synonym generation model. Some visualized examples are shown in Table 2. Tables 3–5 summarize the experimental results on informative queries datasets, offering a comparative analysis of our QA-Attack method against five state-of-the-art QA baselines. For boolean queries, we present the attacking results on the BoolQ dataset in Table 6. Besides, we provide code for the reproducibility of our experiments.[2] The metrics used in our experiment are:

- **F1**: The F1 score balances precision and recall, providing a nuanced view of how much the attacked answers match reference answers.
- **ROUGE and BLEU**: A higher BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) score in context indicates that the adversarial context retains more of the exact phrasing, contributing to better linguistic fluency and coherence.
- **Exact Match (EM)** Measures the percentage of model predictions that exactly match the correct answers in both content and format.
- **Similarity (SIM)**: Evaluates the semantic similarity between original and adversarial context using BERT (Devlin et al., 2019) embeddings. (Note: In our following experiments, EM and SIM are not only measured answers but also reflect the quality of the generated context in Section 4.5.3).
- **Modification Rate (Mod)**: Mod measures the proportion of altered tokens in the text. This metric considers each instance of replacement, insertion, or deletion as a single token modification.
- **Grammar Error (GErr)**: GErr measures the increase in grammatical inaccuracies within successful adversarial examples relative to the original text. This measurement employs LanguageTool (Naber, 2003) to enumerate grammatical errors.
- **Perplexity (PPL)**: PPL serves as an indicator of linguistic fluency in adversarial examples (Kann et al., 2018; Zang et al., 2020). The perplexity calculation utilizes a GPT-2 model with a restricted vocabulary (Radford et al., 2019).

It is important to note that the evaluation metrics (SIM, Mod, GErr, and PPL) employed in this work are specifically designed to quantify the degree of semantic and linguistic divergence between original and adversarial inputs from a computational model perspective. These metrics are used to assess attack effectiveness in disrupting model predictions, rather than directly measuring human perceptibility. We do not assume that lower SIM or GErr/PPL values necessarily correspond to greater detectability by human readers. While human imperceptibility is a focus of other NLP attacks, such as summarization, it is not a focus in our research for QA attacks, due to the different types of application scenarios. It is supported in prior studies that attacks can be evaluated

[1] https://github.com/google-research/bert/?tab=readme-ov-file.
[2] Our code is available at https://github.com/UTSJiyaoLi/QA-Attack.

Table 3
Comparative analysis of QA-Attack and baseline models on T5. Drops of BLEU and ROUGE scores (uni-gram) on contexts are reported in the table, with higher values indicating better performance. For F1, EM, and SIM (i.e., similarity) metrics on answers, lower values indicate better performance.

Datasets	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
SQuAD 1.1	TASA (Cao et al., 2022)	9.21	7.49	89.12	82.88	6.38
	TMYC (Wallace et al., 2019b)	7.28	8.21	81.91	78.72	8.22
	RobustQA (Yasunaga et al., 2018)	5.89	7.52	84.23	77.41	6.03
	TextFooler (Jin et al., 2020)	10.6	10.49	83.11	76.05	6.29
	T3 (Wang et al., 2020)	5.41	6.29	86.83	73.82	7.23
	PIA (Parry et al., 2024)	7.32	8.33	82.37	79.03	7.33
	LLM-Attack (Wang et al., 2023)	8.98	9.54	85.11	75.12	6.77
	QA-Attack (ours)	**4.67**	**5.68**	**90.51**	**84.11**	**5.91**
SQuAD V2.0	TASA (Cao et al., 2022)	20.09	19.31	70.21	76.06	7.29
	TMYC (Wallace et al., 2019b)	17.23	20.68	65.19	69.82	9.05
	RobustQA (Yasunaga et al., 2018)	16.37	18.73	67.71	63.19	8.14
	TextFooler (Jin et al., 2020)	21.69	24.5	65.33	65.01	9.32
	T3 (Wang et al., 2020)	11.19	19.68	69.71	73.53	8.82
	PIA (Parry et al., 2024)	12.58	17.46	71.48	75.46	8.11
	LLM-Attack (Wang et al., 2023)	13.82	19.63	69.01	74.21	8.44
	QA-Attack (ours)	**9.13**	**15.41**	**72.76**	**77.28**	**6.33**
Narrative QA	TASA (Cao et al., 2022)	11.79	15.25	68.11	70.36	6.11
	TMYC (Wallace et al., 2019b)	12.73	9.32	65.91	67.22	7.61
	RobustQA (Yasunaga et al., 2018)	10.01	13.91	67.19	64.11	6.81
	TextFooler (Jin et al., 2020)	14.72	18.61	63.85	62.82	11.74
	T3 (Wang et al., 2020)	11.74	11.37	62.34	60.17	6.28
	PIA (Parry et al., 2024)	11.65	14.81	66.67	70.51	6.71
	LLM-Attack (Wang et al., 2023)	12.23	15.72	63.71	69.04	8.91
	QA-Attack (ours)	**5.61**	**7.23**	**69.18**	**75.73**	**5.23**
NewsQA	TASA (Cao et al., 2022)	8.56	29.44	77.28	69.44	12.28
	TMYC (Wallace et al., 2019b)	6.12	31.23	77.96	72.49	10.32
	RobustQA (Yasunaga et al., 2018)	5.12	29.48	78.72	79.82	10.84
	TextFooler (Jin et al., 2020)	9.01	30.86	74.21	57.44	27.91
	T3 (Wang et al., 2020)	6.21	28.52	75.22	72.56	14.27
	PIA (Parry et al., 2024)	7.41	28.77	75.18	74.91	14.41
	LLM-Attack (Wang et al., 2023)	8.11	29.15	74.33	69.92	11.37
	QA-Attack (ours)	**3.61**	**24.42**	**78.85**	**82.83**	**8.92**

without human evaluations (Cheng et al., 2020; Li et al., 2019; Madry et al., 2018).

4.4. Experiment analysis

Our experimental results in Tables 3–5 demonstrate that QA-Attack consistently outperforms baseline methods across all informative datasets. As shown in Table 6, our method achieves superior performance on the boolean dataset, surpassing all baseline approaches in degrading victim models' accuracy (note that TASA is designed only for informative queries; it is incompatible with boolean query attacks). For informative queries, comparing performance on attacking LongT5 with SQuAD 1.1 and NarrativeQA datasets (representing shortest and longest contexts) in Table 5, we observe that while F1 and EM scores decrease for longer contexts, QA-Attack maintains superiority over baselines. This indicates our approach's robustness and adaptability to varying context lengths, particularly in long texts. The improved performance in longer contexts suggests our HRF approach effectively identifies and targets vulnerable tokens. Regarding semantic consistency, QA-Attack achieves lower similarity scores compared to baseline methods, indicating that the answers generated after the attack deviate more in meaning from the ground truth responses. In addition, as shown in Table 7, these attacks reveal heightened vulnerabilities when applied to specialized medical and financial data inputs. Notably, QA-Attack achieves superior performance on both the EMRQA and FinQA datasets, consistently outperforming all baselines across the informative and boolean answer types, which are categorized based on our task-specific definition.

Additionally, the quality of the generated adversarial samples is evident from the ROUGE and BLEU scores. Our method consistently achieves higher ROUGE and BLEU scores compared to the baselines, which suggests that the adversarial examples generated by QA-Attack are not only effective in terms of altering the model's output but also maintain a high degree of contextual and linguistic coherence. This is largely due to our synonym selection method, which ensures the replacements are contextually appropriate and semantically relevant. Moreover, the token-level replacement strategy, which only modifies fewer words (typically five in the base setting), further ensures that the adversarial examples remain similar to the original context while fooling the model.

4.5. Ablation and hyperparameters studies

To comprehensively validate the efficacy of the proposed QA-Attack method, this section conducts a detailed ablation study, dissecting each component to assess its individual impact and overall contribution to the method's performance.

4.5.1. Effectiveness of hybrid ranking fusion on multiple question types

We test how HRF, ABR, and RBR methods perform across different top_k values on the SQuAD and BoolQ datasets, with d remaining, shown in Fig. 2. HRF consistently outperforms ABR and RBR for all top_k values on both datasets. This suggests that combining attention-based and removal-based ranking in HRF is more effective at generating robust adversarial examples than using either method alone. The graph also shows that as top_k increases, all methods improve, indicating that higher top_k values help identify vulnerable tokens better and lead to more effective attacks.

Despite the better performance at higher top_k values, the study uses $top_k = 5$ as a base setting. This choice balances effectiveness with minimal text modification, ensuring that adversarial examples remain close

Table 4
Comparative analysis of QA-Attack and baseline models on $\text{Bert}_{\text{base}}$. Drops of BLEU and ROUGE scores (uni-gram) on contexts are reported in the table, with higher values indicating better performance. For F1, EM, and SIM (i.e., similarity) metrics on answers, lower values indicate better performance.

Datasets	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
SQuAD 1.1	TASA (Cao et al., 2022)	15.27	34.33	82.87	67.22	8.19
	TMYC (Wallace et al., 2019b)	12.89	28.63	81.51	76.39	10.24
	RobustQA (Yasunaga et al., 2018)	15.72	25.38	79.28	73.27	15.81
	TextFooler (Jin et al., 2020)	23.04	37.28	67.28	49.49	14.11
	T3 (Wang et al., 2020)	8.79	16.11	57.19	63.81	16.92
	PIA (Parry et al., 2024)	27.12	40.56	60.03	42.17	17.94
	LLM-Attack (Wang et al., 2023)	8.23	19.34	86.03	72.81	8.91
	QA-Attack (ours)	**6.42**	**13.31**	**91.22**	**77.16**	**7.43**
SQuAD V2.0	TASA (Cao et al., 2022)	31.22	28.9	77.06	69.05	8.22
	TMYC (Wallace et al., 2019b)	29.38	27.77	73.81	67.23	10.34
	RobustQA (Yasunaga et al., 2018)	27.64	31.82	75.67	71.42	11.23
	TextFooler (Jin et al., 2020)	36.8	29.49	67.14	62.67	13.28
	T3 (Wang et al., 2020)	26.16	27.47	74.94	70.14	7.24
	PIA (Parry et al., 2024)	39.62	34.21	60.81	58.06	14.19
	LLM-Attack (Wang et al., 2023)	25.21	25.88	77.06	71.77	7.31
	QA-Attack (ours)	**22.18**	**21.5**	**80.12**	**75.23**	**4.11**
Narrative QA	TASA (Cao et al., 2022)	12.11	14.51	61.15	63.04	7.32
	TMYC (Wallace et al., 2019b)	8.41	10.23	52.89	69.82	10.72
	RobustQA (Yasunaga et al., 2018)	7.24	9.43	63.81	67.43	9.53
	TextFooler (Jin et al., 2020)	13.74	18.79	56.11	56.82	14.21
	T3 (Wang et al., 2020)	8.49	15.35	65.48	67.09	7.83
	PIA (Parry et al., 2024)	15.33	20.22	53.61	55.04	17.41
	LLM-Attack (Wang et al., 2023)	6.41	11.02	66.42	66.04	6.42
	QA-Attack (ours)	**3.86**	**9.34**	**69.44**	**71.15**	**5.61**
NewsQA	TASA (Cao et al., 2022)	16.85	20.95	68.74	69.12	15.22
	TMYC (Wallace et al., 2019b)	15.86	31.23	77.96	72.49	10.31
	RobustQA (Yasunaga et al., 2018)	17.72	29.48	79.62	67.33	10.84
	TextFooler (Jin et al., 2020)	24.13	22.63	59.17	61.22	31.07
	T3 (Wang et al., 2020)	21.22	22.57	65.14	67.11	18.27
	PIA (Parry et al., 2024)	28.65	33.21	58.77	59.19	33.61
	LLM-Attack (Wang et al., 2023)	18.45	26.51	70.07	69.51	10.12
	QA-Attack (ours)	**14.91**	**20.20**	**80.71**	**74.87**	**9.22**

to the original context while still being effective. The consistent trend across both SQuAD and BoolQ datasets demonstrates that HRF's superior performance holds for different question types, showing its versatility in attacking various question-answering models. This analysis highlights the practical effectiveness of the HRF method and its ability to generate impactful adversarial examples across different QA tasks.

To further examine the relationship between ABR and RBR within HRF, we evaluate various weighting ratios for each method's contribution to the final HRF score. Specifically, we assign varying distribution percentages to ABR and RBR before combining their scores, allowing us to assess the impact of each method on overall performance. As shown in Fig. 3, the x-axis denotes the ABR:RBR weight ratio, while the y-axis reflects the corresponding EM scores. The results reveal that increasing ABR's weight leads to only a slight increase in EM, whereas increasing RBR's weight results in a more significant degradation in performance (much higher EM scores). This trend suggests that ABR contributes more to identifying impactful candidates and plays a more critical role than RBR within the HRF method.

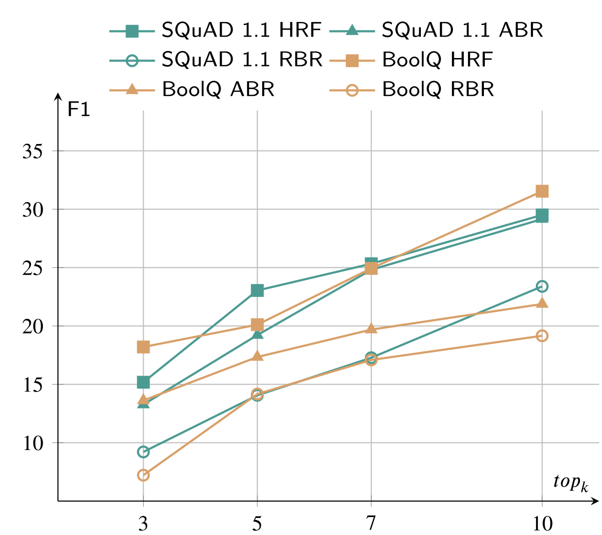


Fig. 2. F1 score analysis for HRF, ABR, and RBR variants of QA-Attack using different top_k values, tested on datasets SQuAD 1.1 and BoolQ.

4.5.2. Effectiveness of synonym selection

To evaluate our Synonyms Selection approach, we conduct comparisons in two aspects. We first compare our BERT-based synonym generation against two alternative methods: WordNet (Miller, 1992), an online database that contains sets of synonyms, and HowNet (Dong & Dong, 2003), which produces semantically similar words using its network structure. Using the base configuration, we evaluate the EM scores when attacking T5 and $\text{BERT}_{\text{base}}$ models across three datasets: SQuAD 1.1, NarrativeQA, and BoolQ. The results in Table 8 demonstrate that our QA-Attack with $\text{BERT}_{\text{base}}$ consistently achieved superior performance compared to other methods across all datasets and victim models.

On the other hand, we also examine the impact of parameter d in Synonym Selection, which determines the number of synonyms obtained from the Masked Language Model (MLM). Table 9 illustrates that as d increases from 1 to 3, F1 scores consistently decrease across all datasets, indicating improved attack performance. This trend suggests that a more aggressive setting (higher d) is more effective in compromising model accuracy across various datasets.

Table 5
Comparative analysis of QA-Attack and baseline models on LongT5. Drops of BLEU and ROUGE scores (uni-gram) on contexts are reported in the table, with higher values indicating better performance. For F1, EM, and SIM (i.e., similarity) metrics on answers, lower values indicate better performance.

Datasets	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
SQuAD 1.1	TASA (Cao et al., 2022)	10.61	22.45	80.67	70.41	11.88
	TMYC (Wallace et al., 2019b)	12.43	29.81	75.37	63.83	13.22
	RobustQA (Yasunaga et al., 2018)	17.22	31.11	73.11	68.29	17.64
	TextFooler (Jin et al., 2020)	35.31	44.09	57.77	49.49	25.33
	T3 (Wang et al., 2020)	9.33	24.52	49.23	60.33	20.87
	PIA (Parry et al., 2024)	8.92	20.11	82.91	70.03	10.32
	LLM-Attack (Wang et al., 2023)	9.31	21.03	81.67	68.82	10.73
	QA-Attack (ours)	**7.38**	**18.78**	**84.22**	**72.67**	**9.67**
SQuAD V2.0	TASA (Cao et al., 2022)	30.71	30.11	64.71	67.28	9.32
	TMYC (Wallace et al., 2019b)	34.11	33.88	64.21	65.11	14.82
	RobustQA (Yasunaga et al., 2018)	29.01	39.59	62.91	68.22	13.09
	TextFooler (Jin et al., 2020)	38.25	34.67	60.47	64.16	15.44
	T3 (Wang et al., 2020)	30.44	30.13	65.81	63.72	8.29
	PIA (Parry et al., 2024)	28.13	27.71	75.62	69.83	7.21
	LLM-Attack (Wang et al., 2023)	29.41	28.44	74.19	68.11	7.64
	QA-Attack (ours)	**27.11**	**24.73**	**77.37**	**70.32**	**5.29**
Narrative QA	TASA (Cao et al., 2022)	8.22	10.67	69.83	65.77	9.53
	TMYC (Wallace et al., 2019b)	9.36	11.33	63.15	64.27	14.72
	RobustQA (Yasunaga et al., 2018)	15.83	12.03	64.28	63.12	12.77
	TextFooler (Jin et al., 2020)	12.77	14.82	62.99	54.21	17.33
	T3 (Wang et al., 2020)	8.38	8.26	63.92	66.32	8.92
	PIA (Parry et al., 2024)	6.31	6.93	68.02	66.91	8.23
	LLM-Attack (Wang et al., 2023)	6.82	7.52	67.11	65.74	8.51
	QA-Attack (ours)	**4.62**	**5.33**	**70.33**	**68.32**	**7.44**
NewsQA	TASA (Cao et al., 2022)	16.85	24.54	64.83	66.81	14.82
	TMYC (Wallace et al., 2019b)	19.28	29.01	62.88	68.67	11.43
	RobustQA (Yasunaga et al., 2018)	17.23	27.42	58.32	57.22	13.37
	TextFooler (Jin et al., 2020)	27.22	26.39	53.33	53.01	25.82
	T3 (Wang et al., 2020)	17.83	25.87	63.25	65.43	19.27
	PIA (Parry et al., 2024)	16.18	25.91	66.81	67.11	11.32
	LLM-Attack (Wang et al., 2023)	16.91	26.41	65.42	65.02	11.78
	QA-Attack (ours)	**15.32**	**24.12**	**68.23**	**70.55**	**10.48**

Table 6
Attack performance comparison on baseline models using the BoolQ dataset, with top results highlighted in bold. Note that TASA (Cao et al., 2022) is not applicable to boolean questions.

Victim Models	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
T5	TASA (Cao et al., 2022)	N/A	N/A	N/A	N/A	N/A
	TMYC (Wallace et al., 2019b)	17.43	19.36	82.09	77.23	21.83
	RobustQA (Yasunaga et al., 2018)	14.33	18.92	79.15	80.33	13.22
	TextFooler (Jin et al., 2020)	20.11	19.07	80.91	83.25	33.82
	T3 (Wang et al., 2020)	15.16	14.74	71.32	68.79	15.82
	PIA (Parry et al., 2024)	11.23	15.11	85.02	81.21	12.94
	LLM-Attack (Wang et al., 2023)	11.91	15.77	84.47	79.87	13.21
	QA-Attack (ours)	**8.64**	**13.9**	**87.31**	**86.57**	**11.42**
$\text{Bert}_{\text{base}}$	TASA (Cao et al., 2022)	N/A	N/A	N/A	N/A	N/A
	TMYC (Wallace et al., 2019b)	21.35	13.28	63.21	70.57	7.34
	RobustQA (Yasunaga et al., 2018)	24.81	9.21	69.22	76.01	6.67
	TextFooler (Jin et al., 2020)	33.02	11.57	65.11	67.81	8.17
	T3 (Wang et al., 2020)	22.06	11.02	76.17	74.62	6.23
	PIA (Parry et al., 2024)	20.13	8.93	74.81	76.31	5.92
	LLM-Attack (Wang et al., 2023)	21.11	9.52	73.92	74.21	6.24
	QA-Attack (ours)	**18.39**	**6.51**	**77.21**	**78.11**	**4.66**
LongT5	TASA (Cao et al., 2022)	N/A	N/A	N/A	N/A	N/A
	TMYC (Wallace et al., 2019b)	29.77	9.82	67.04	73.22	7.43
	RobustQA (Yasunaga et al., 2018)	24.56	8.21	70.49	71.83	9.33
	TextFooler (Jin et al., 2020)	33.02	11.57	65.11	67.81	8.17
	T3 (Wang et al., 2020)	22.06	11.02	76.17	74.62	6.23
	PIA (Parry et al., 2024)	20.23	7.44	74.51	76.42	5.87
	LLM-Attack (Wang et al., 2023)	21.11	8.32	73.23	74.31	6.02
	QA-Attack (ours)	**17.67**	**7.03**	**78.67**	**77.54**	**4.37**

Table 7
Comparative analysis of eight attack methods targeting the T5 model across EMRQA and FinQA datasets, categorized by answer type. Results are stratified into boolean and informative response subsets to illustrate performance variations across different task types.

Answer Type	Datasets	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
Informative	EMRQA	TASA (Cao et al., 2022)	12.25	12.25	61.37	61.54	8.16
		TMYC (Wallace et al., 2019b)	11.38	11.38	55.27	59.18	8.62
		RobustQA (Yasunaga et al., 2018)	10.27	10.27	57.70	58.49	7.04
		TextFooler (Jin et al., 2020)	12.22	12.22	54.33	61.27	8.44
		T3 (Wang et al., 2020)	10.73	10.73	56.49	61.89	7.13
		PIA (Parry et al., 2024)	11.35	11.35	53.96	54.94	7.28
		LLM-Attack (Wang et al., 2023)	11.27	11.27	62.53	55.81	7.62
		QA-Attack (ours)	**8.79**	**8.79**	**64.87**	**62.24**	**6.73**
	FinQA	TASA (Cao et al., 2022)	11.94	11.94	57.56	63.93	7.95
		TMYC (Wallace et al., 2019b)	10.77	10.77	57.89	60.14	8.37
		RobustQA (Yasunaga et al., 2018)	9.91	9.91	60.10	59.83	6.94
		TextFooler (Jin et al., 2020)	11.88	11.88	55.02	59.65	8.79
		T3 (Wang et al., 2020)	9.54	9.54	63.78	55.42	7.30
		PIA (Parry et al., 2024)	10.83	10.83	58.26	57.41	7.23
		LLM-Attack (Wang et al., 2023)	10.81	10.81	62.18	56.32	7.31
		QA-Attack (ours)	**8.59**	**8.59**	**67.42**	**65.11**	**6.61**
Boolean	EMRQA	TASA (Cao et al., 2022)	12.14	12.14	60.64	64.29	7.82
		TMYC (Wallace et al., 2019b)	11.22	11.22	56.17	57.99	8.73
		RobustQA (Yasunaga et al., 2018)	10.14	10.14	61.16	61.16	7.28
		TextFooler (Jin et al., 2020)	12.08	12.08	62.93	57.39	8.87
		T3 (Wang et al., 2020)	9.80	9.80	64.60	62.10	7.05
		PIA (Parry et al., 2024)	10.92	10.92	55.30	58.69	7.50
		LLM-Attack (Wang et al., 2023)	10.74	10.74	60.81	63.54	7.18
		QA-Attack (ours)	**8.91**	**8.91**	**64.98**	**66.92**	**6.78**
	FinQA	TASA (Cao et al., 2022)	12.01	12.01	60.69	54.96	7.91
		TMYC (Wallace et al., 2019b)	10.81	10.81	62.92	63.64	8.66
		RobustQA (Yasunaga et al., 2018)	9.95	9.95	64.15	58.51	7.19
		TextFooler (Jin et al., 2020)	11.93	11.93	60.90	55.68	8.54
		T3 (Wang et al., 2020)	9.36	9.36	60.86	59.71	7.21
		PIA (Parry et al., 2024)	10.91	10.91	60.14	59.15	7.32
		LLM-Attack (Wang et al., 2023)	10.89	10.89	54.67	56.80	7.44
		QA-Attack (ours)	**8.63**	**8.63**	**67.31**	**64.58**	**6.69**

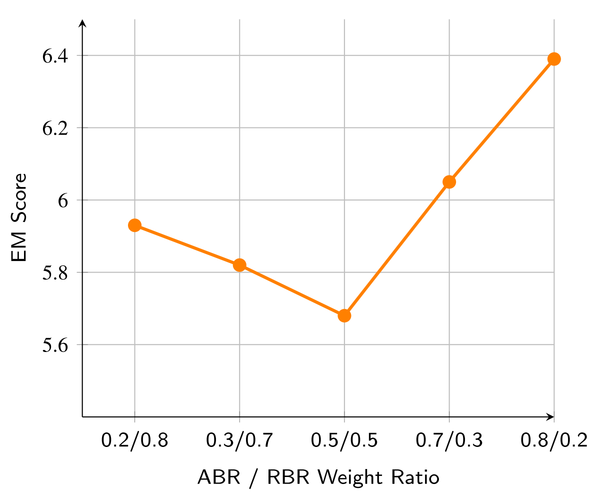


Fig. 3. The impact of HRF weight distribution (ABR/RBR) on EM score using the SQuAD 1.1 dataset. A lower EM score means better attack performance.

Table 8
EM scores for attacks on T5 and $BERT_{base}$ models using three distinct synonym generation methods. Lower scores indicate more effective attacks.

Methods	Victim models	Datasets		
		SQuAD 1.1	NarrativeQA	BoolQ
HowNet	T5	14.22	7.25	29.08
	$BERT_{base}$	7.66	4.52	26.91
WordNet	T5	5.31	3.99	21.63
	$BERT_{base}$	7.23	5.67	19.35
$BERT_{base}$ (ours)	T5	**4.67**	**5.61**	**8.64**
	$BERT_{base}$	**6.42**	**3.86**	**18.39**

Table 9
F1 scores demonstrating QA-Attack's performance across five datasets under different d values (i.e., number of synonym candidates for substitutions).

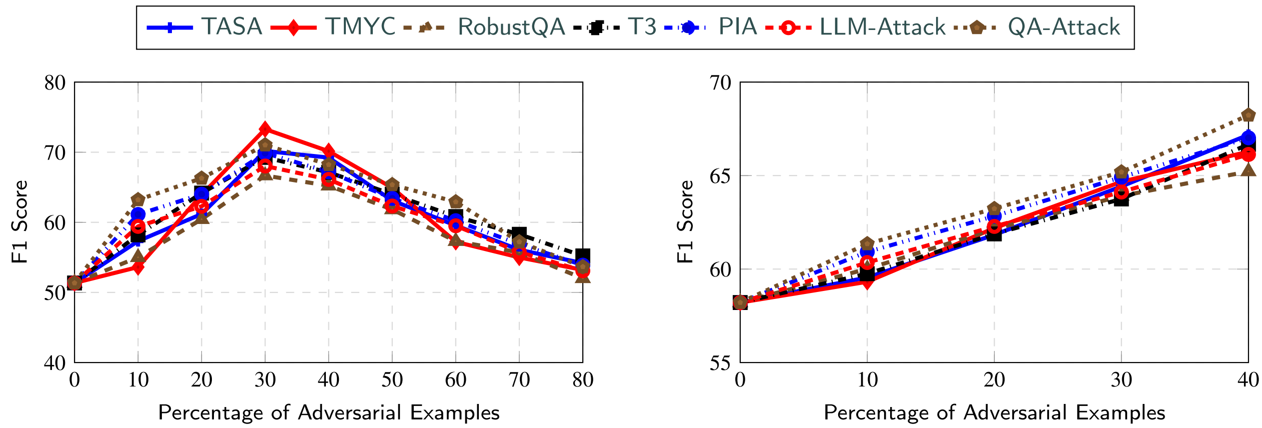
	SQuAD 1.1	SQuAD V2.0	BoolQ	Narrative QA	NewQA
$d = 1$	8.52	14.72	19.22	7.63	10.66
$d = 2$	4.67	9.13	15.16	5.61	3.61
$d = 3$	2.17	7.26	11.43	3.71	3.27

4.5.3. Texual quality of word candidates

In our ablation study, detailed in Table 10, we investigate the quality of adversarial examples generated by various attack methods on the T5 model using the SQuAD 1.1 dataset. We evaluate our word replacement technique with encoder-decoder candidate generation (T3), as well as sentence-level modification methods (TASA, TMYC). The results indicate that our word-level synonym selection approach outperformed all other baselines. Notably, our word-level attack maintains a lower grammar error rate and higher linguistic fluency than alternative methods. Although RobustQA employs the same synonym selection strategy, it requires more word modifications to successfully attack the model and tends to produce more adventurous alterations.

4.6. Platform and efficiency analysis

In this section, we evaluate QA-Attack's computational efficiency under base settings. We measure efficiency using time consumption per sample, expressed in seconds, where a lower value indicates



(a) F1 score trends when retraining with varying sizes of adversarial examples generated by different methods.

(b) F1 scores of attacking T5 models retrained with increasing proportions of adversarial examples.

Fig. 4. F1 scores of T5 model on SQuAD 1.1 dataset showing (a) performance after retraining with varying proportions of adversarial examples from multiple generation methods, and (b) robustness against attacks on the retrained model under different scenarios.

Table 10
Performance metrics for different word candidate selection strategies against the T5 model on the SQuAD 1.1 dataset.

Methods	EM↓	SIM↑	Mod↓	PPL↓	GErr↓
TASA (Cao et al., 2022)	9.21	6.38	8.15	143	0.13
TMYC (Wallace et al., 2019b)	7.28	8.22	9.21	151	0.14
RobustQA (Yasunaga et al., 2018)	5.89	6.03	8.35	147	0.15
T3 (Wang et al., 2020)	6.23	7.23	7.93	133	0.13
TextFooler (Jin et al., 2020)	10.60	6.29	8.17	136	0.14
PIA (Parry et al., 2024)	6.21	6.12	7.85	130	0.13
LLM-Attack (Wang et al., 2023)	6.48	6.04	8.01	132	0.13
QA-Attack (ours)	**5.68**	**5.91**	**7.24**	**125**	**0.12**

Table 11
Time consumption (seconds per sample) for various methods and datasets. A lower value indicates better performance.

	Narrative QA	SQuAD 1.1	SQuAD V2.0	NewsQA	BoolQ
TASA (Cao et al., 2022)	28.77	15.82	18.25	10.72	N/A
TMYC (Wallace et al., 2019b)	25.61	12.75	16.33	9.21	7.42
RobustQA (Yasunaga et al., 2018)	25.82	24.46	22.15	12.81	15.82
T3 (Wang et al., 2020)	26.52	21.37	28.38	14.74	7.93
PIA (Parry et al., 2024)	24.81	12.35	17.67	9.28	7.53
LLM-Attack (Wang et al., 2023)	27.33	14.87	13.09	11.21	7.98
QA-Attack (ours)	**23.51**	**10.61**	**12.38**	**8.32**	**7.22**

superior performance. As shown in Table 11, the outcomes reveal that QA-Attack exhibits remarkable time efficiency, consistently outperforming baseline methods across both long-text (NarrativeQA) and short-text (SQuAD 1.1) datasets. This superior performance can be attributed to QA-Attack's innovative Hybrid Ranking Fusion (HRF) strategy, which effectively identifies vulnerable words within the text, significantly enhancing the speed of the attack process.

4.7. Adversarial retraining

In this section, we investigate QA-Attack's potential for enhancing downstream models' accuracy. We employ QA-Attack to generate adversarial examples from SQuAD 1.1 training sets and incorporate them as supplementary training data. We reconstruct the training set with varying proportions of adversarial examples added to the raw training set. The retraining process with this augmented data aims to examine how test accuracy changes in response to the inclusion of adversarial examples. As illustrated in Fig. 4(a), re-training with adversarial examples slightly improves model performance when less than 30 % of the training data consists of adversaries. However, performance decreases when the proportion of adversaries exceeds 30 %. This finding indicates that the optimal ratio of adversarial examples in training data needs to be determined empirically, which aligns with conclusions from previous attacking methods. To evaluate how re-training helps defend against adversarial attacks, we analyze the robustness of T5 models trained with varying proportions of adversarial examples (0 %, 10 %, 20 %, 30 %, 40 %) from different attack methods, as shown in Fig. 4(b). A lower F1 score indicates higher model susceptibility to adversarial attacks. It demonstrates that incorporating adversarial examples during training consistently improves model robustness, as evidenced by increasing F1 scores across all attack methods. Notably, QA-Attack emerges as the most effective approach, consistently outperforming other methods, with its advantage becoming particularly pronounced at higher percentages of adversarial training data.

4.8. Attacking models with defense mechanism

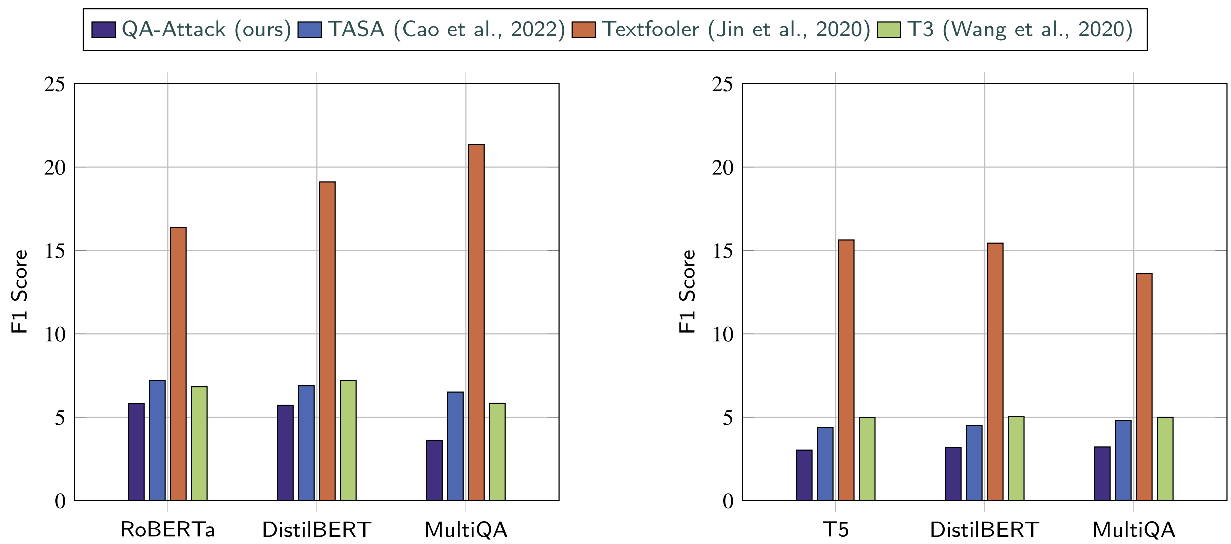
Defending NLP models against adversarial attacks is crucial for maintaining the reliability of language processing systems in real-world applications (Goyal et al., 2023). To further analyze how attacks are performed under defense systems, we deploy two distinct defense mechanisms to investigate our attack performance under defense systems. The first is Frequency-Guided Word Substitutions (FGWS) approach (Mozes et al., 2021), which excels at detecting adversarial examples. The second is Random Masking Training (RanMASK) (Zeng et al., 2023), a technique that enhances model robustness through specialized training procedures. We perform the adversarial attack on T5 on datasets SQuAD 1.1, NarrativeQA, and BoolQ; the results are presented in Table 12. The results show that QA-Attack demonstrates superior adversarial robustness across multiple benchmark datasets, consistently outperforming existing methods against state-of-the-art defenses.

4.9. Transferability of attacks

To evaluate the transferability of our method, we conduct cross-model attacks using adversarial examples generated from two different models: T5 and RoBERTa. Adversarial examples crafted against T5 are evaluated on RoBERTa, DistilBERT, and MultiQA, while those generated against RoBERTa are transferred to T5, DistilBERT, and MultiQA. Fig. 5 presents our transferability results. Both subfigures 5(a) and (b) demonstrate that adversarial examples generated by QA Attack, whether crafted against T5 or RoBERTa, consistently degrade the performance of other QA models. QA-Attack exhibits superior transferability across

Table 12
Effectiveness of defense mechanisms (FGWS (Mozes et al., 2021) and RanMASK (Zeng et al., 2023)) against QA-Attack: EM scores of T5 model output answers across SQuAD 1.1, NarrativeQA, and BoolQ datasets. Lower scores indicate higher attack success against defenses.

Datasets	Defense	TASA	RobustQA	TMYC	T3	PIA	LLM-Attack	QA-Attack
SQuAD 1.1	FGWS (Mozes et al., 2021)	34.71	39.42	28.51	24.11	22.63	23.21	**21.03**
	RanMASK (Zeng et al., 2023)	32.17	39.78	44.81	41.09	31.28	32.02	**30.26**
Narrative QA	FGWS (Mozes et al., 2021)	49.28	44.62	**37.21**	45.17	38.91	39.54	38.33
	RanMASK (Zeng et al., 2023)	38.41	37.14	41.62	43.81	35.81	36.42	**34.47**
BoolQ	FGWS (Mozes et al., 2021)	45.71	47.37	38.97	45.33	39.11	39.94	**38.34**
	RanMASK (Zeng et al., 2023)	41.63	42.88	47.25	42.17	41.12	41.66	**40.51**



(a) Transfer attack with samples generated for T5

(b) Transfer attack with samples generated for RoBERTa

Fig. 5. F1 scores for transfer attacks on QA models using adversarial samples generated for T5 or RoBERTa. Lower values indicate better performance. (a) Transfer attack with samples generated for T5, (b) Transfer attack with samples generated for RoBERTa.

both source model configurations compared to baseline methods (TASA, TextFooler, and T3) on both NarrativeQA and BoolQ datasets. These results validate that QA-Attack generates adversarial perturbations that exploit fundamental vulnerabilities shared across diverse neural QA architectures, demonstrating our approach's robustness and broad applicability.

4.10. Parts of speech preference

To further understand the candidate word distribution of our word-level attack, we examine its attacking preference in terms of Parts of Speech (POS), highlighting vulnerable areas within the input context. We use the Stanford POS tagger (Toutanova et al., 2003) to label each attacked word, categorizing them as *noun*, *verb*, *adjective (Adj.)*, *adverb (Adv.)*, and *others* (e.g., pronoun, preposition, conjunction).

Table 13 illustrates the POS preferences of QA-Attack compared to baseline methods. For *informative queries* on the SQuAD 1.1 dataset, most attacking methods predominantly target *nouns*, which are typically key semantic carriers in questions and contexts. Modifying nouns can directly alter the core meaning of the passage, leading the model to generate incorrect answers. Interestingly, TASA shows a slight preference for *adverbs*, which often modify the certainty or scope of statements, subtly affecting the model's interpretation.

For *boolean queries* on the BoolQ dataset, we observe that some attacks tend to manipulate *adjectives* and *adverbs*. These parts of speech are crucial in yes/no questions because they often determine the polarity, intensity, or qualification of statements (e.g., "always" vs "sometimes", "true" vs "possible"), and small changes can easily flip the correct answer.

Notably, our QA-Attack exhibits a higher tendency to target the "others" category, including pronouns, prepositions, and conjunctions. Although these function words carry relatively less standalone semantic content, modifying them can disrupt the grammatical and sequential structure of the sentence.

Table 13
Part-of-speech preferences in victim word selection across different attack methods (TASA incompatible with Boolean queries).

Datasets	Methods	Noun	Verb	Adj.	Adv.	Others
SQuAD 1.1	TASA (Cao et al., 2022)	N/A	N/A	N/A	N/A	N/A
	TMYC (Wallace et al., 2019b)	**47 %**	*21 %*	11 %	5 %	17 %
	RobustQA (Yasunaga et al., 2018)	**34 %**	13 %	*22 %*	16 %	15 %
	TextFooler (Jin et al., 2020)	**44 %**	13 %	*23 %*	8 %	12 %
	T3 (Wang et al., 2020)	**60 %**	*17 %*	6 %	7 %	10 %
	PIA (Parry et al., 2024)	**41 %**	14 %	*24 %*	11 %	10 %
	LLM-Attack (Wang et al., 2023)	**39 %**	13 %	*22 %*	12 %	14 %
	QA-Attack (ours)	*34 %*	9 %	18 %	3 %	**36 %**
BoolQ	TASA (Cao et al., 2022)	N/A	N/A	N/A	N/A	N/A
	TMYC (Wallace et al., 2019b)	14 %	*19 %*	12 %	**35 %**	20 %
	RobustQA (Yasunaga et al., 2018)	*19 %*	14 %	**27 %**	23 %	17 %
	TextFooler (Jin et al., 2020)	**41 %**	15 %	*27 %*	7 %	10 %
	T3 (Wang et al., 2020)	**42 %**	*13 %*	20 %	16 %	9 %
	PIA (Parry et al., 2024)	**36 %**	14 %	*26 %*	13 %	11 %
	LLM-Attack (Wang et al., 2023)	**34 %**	15 %	*24 %*	14 %	13 %
	QA-Attack (ours)	10 %	19 %	*25 %*	18 %	**28 %**

Table 14
A comparative analysis of attacks on various sizes of BERT models using the SQuAD 1.1 dataset. Lower values indicate better attack performance.

Versions	BERT Tiny	BERT Mini	BERT Medium	BERT Large
Size	L = 2, H = 128	L = 4, H = 256	L = 8, H = 512	L = 24, H = 1024
EM ↓	11.82	13.26	13.31	14.25
F1 ↓	5.67	6.35	6.42	7.24
SIM ↓	6.23	7.12	7.43	8.38

Why certain POS are more effective in misleading the model?. The effectiveness of attacking different POS varies between question types due to the distinct ways models process semantic and syntactic cues. For informative queries (SQuAD 1.1 dataset), answers often depend on accurately identifying nouns or named entities within the context. These are the anchors for understanding "who" or "what" the question refers to, making them high-impact targets. Changing nouns forces the model to either misinterpret the reference or fail to locate the answer span.

In contrast, boolean queries (BoolQ dataset) rely more on assessing logical qualifiers and sentence-level polarity. Therefore, modifying adjectives and adverbs (which influence truth values or intensifiers) greatly affect model predictions. For example, changing "always" to "sometimes" or "likely" to "unlikely" can invert the correct yes/no answer without drastically altering sentence fluency or detectability.

QA-Attack's distinct strategy.. Notably, QA-Attack shows a distinct POS preference with a higher proportion of the "others" category (pronouns, prepositions, conjunctions). Though these function words carry less standalone semantic content, they are crucial for sentence structure and syntactic dependencies. Modifying them subtly disrupts the grammatical and sequential structure of the sentence without noticeably changing its meaning.

These findings suggest that effective attacks are not limited to altering content words (*noun*) to shift meaning but can also exploit syntactic and structural weaknesses (*adv, adj, others*), a strategy that underlies QA-Attack's superior performance across diverse QA tasks.

4.11. Robustness versus the scale of pre-trained models

From the attacking results in Table 4 discussed in Section 4.4, we recognize the limitation of our QA-Attack on $\text{BERT}_{\text{base}}$, with $L = 12$ and $H = 768$, which does not sufficiently support robust experimental outcomes. To address this issue and gain more comprehensive insights, we conducted experiments with four different sizes of BERT (Devlin et al., 2019) models[3]: $\text{BERT}_{\text{tiny}}$, $\text{BERT}_{\text{mini}}$, $\text{BERT}_{\text{medium}}$, and $\text{BERT}_{\text{large}}$. Our findings, detailed in Table 14, demonstrate a positive correlation between model size and experimental robustness. The effectiveness of adversarial attacks decreases as the complexity and capacity of the BERT model increase, suggesting that deeper architectures provide better protection against adversarial perturbations.

5. Conclusion and future work

The robustness of QA models has been increasingly challenged by adversarial attacks. These attacks expose the vulnerabilities of models used in various tasks, including information retrieval, conversational agents, and machine comprehension. To address this, we introduced QA-Attack, which leverages Hybrid Ranking Fusion (HRF) to conduct effective attacks by identifying and modifying the most critical tokens in the input text. Through a combination of attention-based and removal-based ranking strategies, QA-Attack successfully disrupts model predictions while maintaining high levels of semantic and linguistic coherence. Extensive experiments have demonstrated that our method outperforms existing attack techniques regarding attack success, fluency, and consumption across various datasets, confirming its efficacy in undermining the robustness of state-of-the-art QA models.

While adversarial attacks such as QA-Attack expose vulnerabilities in QA systems, they simultaneously provide valuable opportunities to evaluate and enhance model robustness. Moving forward, we plan to focus our research on developing effective defense strategies that can mitigate these identified vulnerabilities. In future work, we intend to expand our investigation beyond the current word-level perturbation constraints by incorporating sentence-level attacks, which will provide deeper insights into the impact of more extensive modifications on QA model performance. Additionally, we plan to extend QA-Attack to handle increasingly complex and diverse QA scenarios, including multiple-choice questions and multi-hop reasoning tasks (Yu et al., 2020). Moreover, we intend to investigate targeted attacks designed to provoke model hallucinations, with the goal of understanding and mitigating the factors that lead models to generate unsupported responses beyond QA scenarios.

CRediT authorship contribution statement

Jiyao Li: Conceptualization, Methodology, Writing – original draft; **Mingze Ni:** Investigation, Visualization; **Yongshun Gong:** Writing – review & editing; **Wei Liu:** Writing - review & editing, Methodology.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

[3] Different sizes of BERT models can be obtained from https://github.com/google-research/bert/

References

Antaki, F., Milad, D., Chia, M. A., Giguère, C.-É., Touma, S., El-Khoury, J., Keane, P. A., & Duval, R. (2024). Capabilities of GPT-4 in ophthalmology: An analysis of model entropy and progress towards human-level medical question answering. *British Journal of Ophthalmology*, *108*(10), 1371–1378.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473

Bongini, P., Becattini, F., & Del Bimbo, A. (2022). Is GPT-3 all you need for visual question answering in cultural heritage? In *European conference on computer vision* (pp. 268–281). Springer.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc. (*vol. 33*).

Cao, Y., Li, D., Fang, M., Zhou, T., Gao, J., Zhan, Y., & Tao, D. (2022). TASA: Deceiving question answering models by twin answer sentences attack. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 11975–11992). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Chen, Q., Zhang, Z., Liu, Z., Zhang, H., Liu, Y., & Huang, M. (2022). FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 60th annual meeting of the association for computational linguistics (ACL)*.

Cheng, M., Yi, J., Chen, P.-Y., Zhang, H., & Hsieh, C.-J. (2020). Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 3601–3608.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 2924–2936). Minneapolis, Minnesota: Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dong, H., Dong, J., Yuan, S., & Guan, Z. (2022). Adversarial attack and defense on natural language processing in deep learning: A survey and perspective. In *International conference on machine learning for cyber security* (pp. 409–424). Springer.

Dong, Z., & Dong, Q. (2003). Hownet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. proceedings. 2003* (pp. 820–824). IEEE.

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 31–36). Melbourne, Australia: Association for Computational Linguistics.

Galassi, A., Lippi, M., & Torroni, P. (2021). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(10), 4291–4308.

Garg, S., & Ramakrishnan, G. (2020). Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6174–6181). Association for Computational Linguistics.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *3rd international conference on learning representations (ICLR)* (p. online).

Goyal, S., Doddapaneni, S., Khapra, M. M., & Ravindran, B. (2023). A survey of adversarial defenses and robustness in nlp. *ACM Computing Surveys*, *55*(14s), 1–39.

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H., & Yang, Y. (2022). LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the association for computational linguistics: NAACL 2022* (pp. 724–736). Seattle, United States: Association for Computational Linguistics.

Hathaliya, J. J., Tanwar, S., & Sharma, P. (2022). Adversarial learning techniques for security and privacy preservation: A comprehensive review. *Security and Privacy*, *5*(3), e209.

Iyyer, M., Wieting, J., Gimpel, K., & Zettlemoyer, L. (2018). Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1875–1885). New Orleans, Louisiana: Association for Computational Linguistics.

Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2021–2031). Copenhagen, Denmark: Association for Computational Linguistics.

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8018–8025.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, *11*(14), 6421.

Kann, K., Rothe, S., & Filippova, K. (2018). Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd conference on computational natural language learning* (pp. 313–323). Brussels, Belgium: Association for Computational Linguistics.

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., & Hajishirzi, H. (2020). UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 1896–1907). Online: Association for Computational Linguistics.

Klein, T., & Nabi, M. (2019). Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds. arXiv preprint arXiv:1911.02365

Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 conference on designing interactive systems* DIS '17 (p. 555-565). New York, NY, USA: Association for Computing Machinery.

Ko, M., Lee, J., Kim, H., Kim, G., & Kang, J. (2020). Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 1109–1121). Online: Association for Computational Linguistics.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., & Grefenstette, E. (2018). The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, *6*, 317–328.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. In *International conference on learning representations* (p. N/A).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880). Online: Association for Computational Linguistics.

Li, D., Zhang, Y., Peng, H., Chen, L., Brockett, C., Sun, M.-T., & Dolan, B. (2021). Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 5053–5069). Online: Association for Computational Linguistics.

Li, J., Ji, S., Du, T., Li, B., & Wang, T. (2019). Textbugger: Generating adversarial text against real-world applications. In *Proceedings 2019 network and distributed system security symposium* NDSS 2019. Internet Society.

Li, J., & Liu, W. (2023). Summarization attack via paraphrasing (student abstract). In *Proceedings of the AAAI conference on artificial intelligence* (pp. 16250–16251). (*vol. 37*).

Li, L., Ma, R., Guo, Q., Xue, X., & Qiu, X. (2020). BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6193–6202). Online: Association for Computational Linguistics.

Liang, B., Li, H., Su, M., Bian, P., Li, X., & Shi, W. (2018). Deep text classification can be fooled. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (p. 4208-4215). International Joint Conferences on Artificial Intelligence Organization.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1412–1421). Lisbon, Portugal: Association for Computational Linguistics.

Lyu, W., Zheng, S., Pang, L., Ling, H., & Chen, C. (2023). Attention-enhancing backdoor attacks against BERT-based models. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 10672–10690). Singapore: Association for Computational Linguistics.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International conference on learning representations*.

Martinez-Gil, J. (2023). A survey on legal question–answering systems. *Computer Science Review*, *48*, 100552.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the 1st international conference on learning representations (ICLR)*.

Miller, G. A. (1992). WordNet: A lexical database for English. In *Speech and natural language: Proceedings of a workshop held at Harriman, New York, February 23–26, 1992* (pp. 259–262).

Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2574–2582).

Mozes, M., Stenetorp, P., Kleinberg, B., & Griffin, L. (2021). Frequency-guided word substitutions for detecting textual adversarial examples. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume* (pp. 171–186). Online: Association for Computational Linguistics.

Naber, D. (2003). A rule-based style and grammar checker. GRIN Verlag.

Ni, M., Wang, C., Zhu, T., Yu, S., & Liu, W. (2022). Attacking neural machine translations via hybrid attention learning. *Machine Learning*, *111*(11), 3977–4002.

Nuruzzaman, M., & Hussain, O. K. (2018). A survey on chatbot implementation in customer service industry through deep neural networks. In *2018 IEEE 15th international conference on e-business engineering (ICEBE)* (pp. 54–61).

Pampari, A., Raghavan, P., Liang, J., & Peng, J. (2018). emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP)* (pp. 2359–2369).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Parry, A., Fröbe, M., MacAvaney, S., Potthast, M., & Hagen, M. (2024). Analyzing adversarial attacks on sequence-to-sequence relevance models. In *European conference on information retrieval* (pp. 286–302). Springer.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 784–789). Melbourne, Australia: Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2383–2392). Austin, Texas: Association for Computational Linguistics.

Ren, S., Deng, Y., He, K., & Che, W. (2019). Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1085–1097). Florence, Italy: Association for Computational Linguistics.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 856–865). Melbourne, Australia: Association for Computational Linguistics.

Rosenthal, S., Bornea, M., & Sil, A. (2021). Are multilingual BERT models robust? A case study on adversarial attacks for multilingual question answering. arXiv preprint arXiv:2104.07646

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th workshop on energy efficient machine learning and cognitive computing*. Vancouver, Canada: NeurIPS.

Shinoda, K., Sugawara, S., & Aizawa, A. (2023). Penalizing confident predictions on largely perturbed inputs does not improve out-of-distribution generalization in question answering. In *Proceedings of the workshop on knowledge augmented methods for NLP (knowledgeNLP) at AAAI 2023*.

Soares, M. A. C., & Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, *32*(6), 635–646.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008–3021.

Sun, H., Zhu, T., Zhang, Z., Jin, D., Xiong, P., & Zhou, W. (2021). Adversarial attacks against deep generative models on data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3367–3388.

Talmor, A., & Berant, J. (2019). MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4911–4921). Florence, Italy: Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology - volume 1, NAACL '03* (p. 173–180). USA: Association for Computational Linguistics.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., & Suleman, K. (2017). NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd workshop on representation learning for NLP* (pp. 191–200). Vancouver, Canada: Association for Computational Linguistics.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019a). Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 2153–2162). Hong Kong, China: Association for Computational Linguistics.

Wallace, E., Rodriguez, P., Feng, S., Yamada, I., & Boyd-Graber, J. (2019b). Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, *7*, 387–401.

Wang, B., Pei, H., Pan, B., Chen, Q., Wang, S., & Li, B. (2020). T3: Tree-autoencoder constrained adversarial text generation for targeted attack. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6134–6150). Online: Association for Computational Linguistics.

Wang, Z. (2022). Modern question answering datasets and benchmarks: A survey. arXiv preprint arXiv:2206.15030

Wang, Z., Wang, W., Chen, Q., Wang, Q., & Nguyen, A. (2023). Generating valid and natural adversarial examples with large language models.

Xiao, T., & Zhu, J. (2023). Introduction to transformers: an NLP perspective.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048–2057). PMLR.

Yang, X., Liu, W., Zhang, S., Liu, W., & Tao, D. (2020). Targeted attention attack on deep learning models in road sign recognition. *IEEE Internet of Things Journal*, *8*(6), 4980–4990.

Yasunaga, M., Kasai, J., & Radev, D. (2018). Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 976–986). New Orleans, Louisiana: Association for Computational Linguistics.

Yigit, G., & Amasyali, M. F. (2024). From text to multimodal: A comprehensive survey of adversarial example generation in question answering systems. *Knowledge and Information Systems*, *66*, 7165–7204.

Yin, Z., Wang, F., Liu, W., & Chawla, S. (2018). Sparse feature attacks in adversarial learning. *IEEE Transactions on Knowledge and Data Engineering*, *30*(6), 1164–1177.

Yu, J., Liu, W., Qiu, S., Su, Q., Wang, K., Quan, X., & Yin, J. (2020). Low-resource generation of multi-hop reasoning questions. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6729–6739). Online: Association for Computational Linguistics.

Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., & Sun, M. (2020). Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6066–6080). Online: Association for Computational Linguistics.

Zeng, J., Zheng, X., Xu, J., Li, L., Yuan, L., & Huang, X. (2023). Certified robustness to text adversarial attacks by randomized [MASK]. *Computational Linguistics*, *49*(2), 395–427.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning* (pp. 11328–11339). PMLR.

Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., & Chua, T.-S. (2021). Retrieving and reading: A comprehensive survey on open-domain question answering. arXiv preprint arXiv:2101.00774

Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics* (pp. 1218–1227). Huhhot, China: Chinese Information Processing Society of China.