Pay Attention to Real World Perturbations! Natural Robustness Evaluation in Machine Reading Comprehension

Anonymous ACL submission

Abstract

As neural language models achieve human-comparable performance on Machine Reading Comprehension (MRC) and see widespread adoption, ensuring their robustness in real-world scenarios has become increasingly important. Current robustness evaluation research, though, primarily develops synthetic perturbation methods, leaving unclear how well they reflect real life. Considering this, we present a framework to automatically examine MRC models on occurring textual perturbations, by replacing paragraph in MRC benchmarks with their counterparts based on available Wikipedia edit history. Such perturbation type is *natural* as its design does not stem from an artificial generative process, inherently distinct from the previously investigated synthetic approaches. In a large-scale study encompassing various model architectures we observe that natural perturbations result in performance degradation in pre-trained encoder langauge models, with errors extending to Flan-T5 and Large Language Models (LLMs). We also show that exposing encoder-only models to naturally perturbed examples during training contributes to handling natural perturbations. This adversarial training approach, however, is not able to promote performance improvement on the majority of synthetic perturbations, indicating that many types of synthetic noise do not actually exist in our collected real-world textual perturbations. We hope this study will inspire future robustness investigation efforts to focus more on natural perturbations, thus deepening our understanding of how models respond to realistic linguistic challenges and providing insights into practical robustness enhancement strategies.

1 Introduction

Transformer-based pre-trained language models demonstrate remarkable efficacy in addressing questions based on a given passage of text, a task commonly referred to as Machine Reading Comprehension (MRC) (Devlin et al., 2019; Brown et al., 2020; He et al., 2021; Wei et al., 2022; Touvron et al., 2023; OpenAI et al., 2024). Despite these advancements, high-performing MRC systems are also known to succeed by relying on *shortcuts* in benchmark datasets rather than truly demonstrating understanding of the passage, thereby lacking robustness to various types of test-time perturbations (Ho et al., 2023; Schlegel et al., 2023; Levy et al., 2023).

Evaluating models' resilience to textual perturbations during inference aids in identifying adversarial instances that highlight their shortcut behavior and provides insights into mitigating these shortcuts. While numerous synthetic perturbation approaches have been explored and reveal the vulnerabilities of MRC models to various linguistic challenges (Ribeiro et al., 2018; Jiang and Bansal, 2019; Welbl et al., 2020; Schlegel et al., 2021; Cao et al., 2022; Tran et al., 2023), a serious concern is that these carefully designed perturbations might not necessarily appear in real-world settings. Consequently, this poses a risk of neglecting the weaknesses of reading comprehension systems to real challenges when deployed in practical scenarios, thus potentially hindering the improvement of their reliability in practical applications.

To counteract this issue, in this paper, we develop a framework to inject textual changes that arise in real-world conditions into MRC datasets and audit how well contemporary language models perform under such perturbations. We deem them as natural because the perturbation process does not involve any artificial manipulation, in line with the definitions by Belinkov and Bisk (2018); Hendrycks et al. (2021); Pedraza et al. (2022); Agarwal et al. (2022). Results of robustness evaluation are therefore more representative of real-world applications. Inspired by Belinkov and Bisk (2018), our approach utilises English Wikipedia revision histories as the source of natural perturbations, given that the differences between revi-

sions authentically capture the textual modifications made by human editors in the real world. By comparing the variances between each adjacent revision, we identify perturbed versions for each Wikipedia reading passage in the original MRC benchmarks (if it exists). Apart from natural perturbations, we also investigate the robustness of MRC models to diverse synthetic perturbation techniques at different levels to discern their differences. All perturbation methods only alter the reading context, while the questions and ground truth answers remain unchanged.

With the established framework, we conduct extensive experiments on two datasets, evaluating twenty-nine models, including nine recently proposed LLMs. Experimental results indicate that natural perturbations encompass rich linguistic variations and can lead to failures in the encoder-only models, while humans are almost undeterred by their presense. Crucially, these errors also transfer to larger and more powerful models, such as `Flan-T5` and state-of-the-art LLMs. On synthetic perturbations, we also observe a lack of robustness across all model architectures, although the realism of thusly generated adversarial examples remains a concern. Using naturally perturbed MRC instances for retraining usually boosts the robustness of encoder-only models against natural perturbations. However, this often comes with a decline in original performance. Retraining also sometimes helps improve robustness to synthetic perturbations, but for the majority of cases, it does not enhance performance on the perturbed test set and even decrease it, likely due to the inherent differences between natural and synthetic perturbations.

2 Related Work

Robustness Evaluation in MRC A typical approach to evaluate the robustness of MRC models is via test-time perturbation. This line of research develops different perturbation methods as attacks, such as adversarial distracting sentence addition (Jia and Liang, 2017; Tran et al., 2023), word substitution (Wu et al., 2021), character swap (Si et al., 2021), entity renaming (Yan et al., 2022) and paraphrasing (Gan and Ng, 2019; Lai et al., 2021; Wu et al., 2023a). Our work also fits within the category of test-time perturbation, but differs from previous works in that we introduce perturbations that naturally occur in real-world scenarios, therefore contributing to a more practical robustness examination. We also experiment with various synthetic perturbations for comparison purposes.

Natural Perturbation for Robustness Assessment Compared with deliberately crafting the perturbed instances, the study of natural perturbation is under-explored. In the computer vision domain, researchers find that real-world clean images without intentional modifications can confuse deep learning models as well, terming them as natural adversarial examples (Hendrycks et al., 2021; Pedraza et al., 2022). Similarly, in the field of Natural language processing (NLP), Belinkov and Bisk (2018) concludes that naturally occurring errors dramatically break machine translation systems. Motivated by these, we attempt to harvest natural perturbations from available Wikipedia revision histories and utilise them to modify the original MRC instances. To the best of our knowledge, we are the first to investigate MRC model robustness under real natural perturbations. Furthermore, it should be noted that the concept of natural perturbed examples in this paper differs from what is defined in previous NLP literature, where the latter measures the extent to which synthetically modified text preserves certain linguistic characteristics such as fluency, coherence, grammaticality and clarity, i.e., its naturalness (Jin et al., 2020; Li et al., 2020; Schlegel et al., 2021; Qi et al., 2021; Wang et al., 2022a; Dyrmishi et al., 2023). Some works also propose that a natural synthetically perturbed sample should be imperceptible to human judges (Li et al., 2020; Garg and Ramakrishnan, 2020) or convey the impression of human authorship (Dyrmishi et al., 2023). However, this proposition remains a subject of debate (Zhao et al., 2018; Wang et al., 2022b; Chen et al., 2022b).

3 Natural and Synthetic Perturbation

In this section, we detail our methodology to create label-preserving stress MRC test sets by introducing real-world occurring noises and artificial perturbations to the context paragraph, respectively. The size of each examined MRC dataset and the correspondingly constructed perturbed test sets are displayed in Appendix A.

3.1 Natural Perturbation

We design a pipeline to automatically construct MRC test sets with noises that occur in real-world settings by leveraging Wikipedia revision histories. Our approach comprises two modules: *candidate*

passage pairs curation and *perturbed test set construction*.

Candidate passage pairs curation. For each English Wikipedia article within the development set of MRC datasets, we systematically extract its entire revision histories and preprocess them, including the removal of markups and the segmentation of content. Subsequently, we obtain the content differences between each current revision and the previous adjacent one, identifying three distinct editing patterns: addition, deletion, and modification[1]. In the case of an edit falling within the modification pattern, we retain the paragraph from the prior version as the *original* and the corresponding one from the current version as the *perturbed*, provided both paragraphs exceed 500 characters[2]. This results in a total of 91,093 pairs of candidate reading passages distributed across 46 articles.

Perturbed test set construction. To generate the naturally perturbed test set, we begin by acquiring all reading passages from the development set of each MRC dataset and identifying their entries in the collection of previously extracted candidate original passages, along with the corresponding perturbed counterparts. Subsequently, for the matched original passages with a single occurrence, we keep them and the corresponding perturbed passages; whereas for those with multiple occurrences, we randomly select one instance for each and extract its perturbed version. After obtaining the perturbed reading passages, we retain only those with at least one question where all annotated ground truth answers (or all plausible answers for the unanswerable question) can still be located within the perturbed context, resulting in the *Perturbed* test set. For the sake of comparison, we also construct an *Original* version of the test set keeping only the original passages and questions corresponding to those that were included in the *Perturbed* version.

3.2 Synthetic Perturbation

To explore the difference between natural and synthetic perturbations, our study incorporates a comprehensive range of synthetic perturbation techniques, spanning various linguistic levels: character, word, sentence and document-level, as detailed in Table 1. While certain character-level and word-level perturbation methods have been investigated across multiple NLP tasks, such as the CharSwapMid for machine translation (Belinkov and Bisk, 2018) and the WDelete for quality estimation (Kanojia et al., 2021; Wu et al., 2023b), none of these has been applied to the contextual paragraph to study the robustness for the task of extractive MRC. We employ methods including WSplit, WSynSub and WInsert (WE) to each sentence in the original reading passage, and then recombine the modified sentences to generate the perturbed version. Conversely, other perturbation approaches are directly executed on the entire paragraph, as implementing them at the sentence-level might result in perturbed text that is even difficult for humans to read and comprehend (Si et al., 2021). The implementation of all character-level and word-level methods is carried out using the NLPAug library (Ma, 2019). Moreover, we set the perturbation rate to 30%, in line with the default settings within the NLPAug library.

To conduct a comprehensive evaluation, we also apply a variation of the sentence-level perturbation method known as AddSent (Jia and Liang, 2017; Chen et al., 2022a; Tran et al., 2023; Levy et al., 2023). Our approach prompts the `GPT-3.5-turbo-0125` model to generate a distractor sentence that shares significant lexical overlap with the question but is not an appropriate answer to it and is also irrelevant to the context. Unlike previous approaches, we do not explicitly require the distractor sentence to be a statement that answers the so-called "almost detail" question (Levy et al., 2023). Instead, the model can generate any type of sentence as long as it satisfies the three criteria stated in the prompt. Subsequently, from the set of generated candidate distractor sentence-question pairs under each reading passage, we select the one with the highest lexical overlap. We finally insert the distractor sentence from the identified pair at the beginning of the original context, as previous literature suggests that prepending results in a larger impact on the performance (Ko et al., 2020; Chen et al., 2022a). For document-level perturbations, we introduce two methods that both leverage the capabilities of the `GPT-3.5-turbo-0125` model as well. The former DocPara attempts to directly paraphrase the entire context paragraph, while the latter Style Transfer, drawing inspiration from (Qi et al., 2021), seeks to transfer the style of the reading passage by rephrasing it using a distinct persona

[1] In Appendix B, the average percentage of editing patterns observed for each Wikipedia article in the investigated MRC development datasets is presented, revealing that "modification" constitutes the predominant editing pattern.

[2] This threshold setting adheres to the methodology employed in the collection of SQuAD 1.1 (Rajpurkar et al., 2016).

discerned based on its topic. The manually constructed prompt for methods AddSent, DocPara and Style Transfer are shown in Appendix C.

For each perturbation method, excluding AddSent, which preserves all ground truth answers in nature, we conduct the same answers-preserving checking as described in Section 3.1 after obtaining the perturbed reading passages to construct the corresponding *Original* and *Perturbed* test set pair for evaluation purposes.

Method	Explanation
	character-level
CharOCR	Replace characters with predefined Optical Character Recognition (OCR) errors.
CharInsert	Inject new characters randomly.
CharSubstitute	Substitute original characters randomly.
CharSwapMid	Swap adjacent characters within words randomly, excluding the first and last character.
CharSwapRand	Swap characters randomly without constraint.
	word-level
WInsert (CWE)	Insert new words to random position according to contextual word embeddings calculation from RoBERTa-base.
WSubstitute (CWE)	Substitute words according to contextual word embeddings calculation from RoBERTa-base (Liu et al., 2019).
WSplit	Split words to two tokens randomly.
WSwap	Swap adjacent words randomly.
WDelete	Delete words randomly.
WCrop	Remove a set of continuous word randomly.
Word Synonym Substitution (WSynSub)	Substitute words with synonyms from large size English PPDB (Pavlick et al., 2015).
WInsert (WE)	Insert new words to random position according to GloVe (Pennington et al., 2014) word embeddings calculation[3].
	sentence-level
AddSent	Add a context-irrelevant distractor sentence with high lexical overlap to the question at the beginning of the context.
	document-level
Document Paraphrasing (DocPara)	Paraphrasing the whole context paragraph directly.
Style Transfer	Rephrase the passage using a distinct persona discerned based on its topic.

Table 1: Various synthetic perturbation approaches.

4 Experiments Setup

4.1 Datasets

We select two widely studied benchmark MRC datasets (License: CC-BY-SA-4.0) for which human performance has been surpassed by state-of-the-art models, due to the fact that their reading passages are sourced from Wikipedia, thereby enabling the utilisation of Wikipedia editing histories to generate the naturally perturbed test set.

SQuAD 1.1 (Rajpurkar et al., 2016): An English reading comprehension dataset with over 100,000 questions created by crowdworkers on a set of Wikipedia article paragraphs. Each question is accompanied by multiple ground truth answers, each of which represents a continuous span from the corresponding reading passage.

SQuAD 2.0 (Rajpurkar et al., 2018): The combination of SQuAD 1.1 with over 50,000 unanswerable questions crafted adversarially by crowdworkers, thus considered to be more challenging.

4.2 Models

Our evaluation study involves multiple contemporary MRC models across three different types: encoder-only, encoder-decoder, and decoder-only. Under the encoder-decoder and decoder-only model evaluation settings, we reframe the extractive MRC as the text generation task based on the given context and question. Access to and experimentation with all models are possible via the use of the HuggingFace's *Transformers* library (Wolf et al., 2020), two 80GB Nvidia A100 GPUs and the OpenAI ChatGPT API.

Encoder-only: We select `BERT` (Devlin et al., 2019) and its various variants for evaluation, including `DistilBERT` (Sanh et al., 2019), `SpanBERT` (Joshi et al., 2020), `RoBERTa` (Liu et al., 2019), `ALBERT` (Lan et al., 2020) and `DeBERTa` (He et al., 2021). Some of these model types also come with different variations, such as size (e.g., *base* and *large* for `RoBERTa`), versions (e.g., *v1* and *v2* for `ALBERT`) and whether the input text is cased or not (e.g., *cased* and *uncased* for `BERT`), all of which are included in the evaluation. We fine-tune these encoder-only pre-trained language models on the training set of our examined MRC datasets and evaluate them on the constructed original and perturbed test sets. Model details and the hyperparameters used in model fine-tuning are shown in Appendix D.

Encoder–Decoder: Instruction finetuning has been demonstrated to be effective in enhancing zero-shot performance of pretrained language models, resulting in the development of Finetuned Language Net (FLAN) (Wei et al., 2022). In this work, we use the instruction-finetuned version of T5 model class, specifically the `Flan-T5` (Chung et al., 2022), available in sizes ranging from *small* (80M), *base* (250M), *large* (780M) to *xl* (3B). During evaluation, we utilise the instruction templates from MRC task collection in open-sourced FLAN repository and report the model performance as the average of those obtained across the employed templates. Refer to Appendix E for various instruction templates used for the evaluation on the test sets with the format as the two examined MRC datasets.

Decoder-only: There is an exponential increase of pre-trained generative LLMs and their fine-tuned chat versions, inspired by the remarkable success of ChatGPT (Bang et al., 2023). Therefore, our experiments incorporate a broad range of recently proposed language model families, including `GPT 3.5 Turbo`, `Llama 2` (Touvron et al., 2023), `Llama 3`, `Mistral` (Jiang et al., 2023), `Falcon` (Almazrouei et al., 2023) and `Gemma` (Mesnard et al., 2024). The zero-shot prompts designed for soliciting responses from them are presented in Appendix F.

4.3 Model Evaluation Metrics

We choose the Exact Match (EM) and (Macro-averaged) F1 score to assess the performance of both encoder-only and encoder-decoder models, as they strictly output the shortest continuous span from the context as the answer (or predict the question as unanswerable) during inference. However, for almost all of the decoder-only models, their outputs are not consistently adhere to the instruction due to their conversational style, rendering EM and F1 metrics unsuitable for evaluation. Consequently, we employ a more lenient metric, namely Inclusion Match (IM), which measures whether the response of the model contains any of the ground truth answers. Furthermore, if the model's output includes phrases such as "I cannot answer this/the question" and "unanswerable"[4], we deem that the model believes the question is not answerable. Model robustness is quantified by measuring the relative variation in performance (as reflected in the F1 or IM) under perturbations.

5 MRC under Natural Perturbation

5.1 Are Encoder-only MRC Models Resilient to Natural Perturbation?

Table 2 presents the relative F1 change for all encoder-only MRC models on the naturally perturbed test set generated based on the SQuAD 1.1 and SQuAD 2.0 development set, respectively. It can be clearly seen from Table 2 that overall, the performance of all the examined models decreases, indicating that ***encoder-only MRC models suffer from natural perturbation***. However, we notice that the performance drop of all models is negligible (the biggest drop is only 3.06%), which suggests that those models also exhibit considerable robustness to natural perturbations.

Dataset	Victim \ Attacker	Nature
SQuAD 1.1	`distilbert-base`	-0.6
	`bert-base-cased`	-0.21
	`bert-base-uncased`	-0.87
	`bert-large-cased`	-0.63
	`bert-large-uncased`	-0.35
	`spanbert-base-cased`	-0.26
	`spanbert-large-cased`	-0.51
	`roberta-base`	-0.61
	`roberta-large`	-0.29
	`albert-base-v1`	-1.0
	`albert-base-v2`	-0.34
	`albert-large-v1`	-0.42
	`albert-large-v2`	-0.8
	`albert-xxlarge-v1`	-0.75
	`albert-xxlarge-v2`	-0.46
	`deberta-large`	-0.52
SQuAD 2.0	`distilbert-base`	$-0.71_{(-2.76/1.71)}$
	`bert-base-cased`	$-0.63_{(-1.84/0.6)}$
	`bert-base-uncased`	$-0.49_{(-1.88/0.94)}$
	`bert-large-cased`	$-0.53_{(-1.61/0.55)}$
	`bert-large-uncased`	$-1.38_{(-2.51/-0.24)}$
	`spanbert-base-cased`	$-1.24_{(-2.66/0.15)}$
	`spanbert-large-cased`	$-1.2_{(-1.9/-0.56)}$
	`roberta-base`	$-0.6_{(-2.09/0.81)}$
	`roberta-large`	$-1.52_{(-2.6/-0.54)}$
	`albert-base-v1`	$-1.07_{(-2.02/-0.22)}$
	`albert-base-v2`	$-1.08_{(-2.03/-0.22)}$
	`albert-large-v1`	$-0.41_{(-1.42/0.52)}$
	`albert-large-v2`	$-0.69_{(-1.66/0.22)}$
	`albert-xxlarge-v1`	$-1.23_{(-3.06/0.49)}$
	`albert-xxlarge-v2`	$-1.28_{(-3.02/0.36)}$
	`deberta-large`	$-1.05_{(-2.2/0.0)}$

Table 2: Relative F1 change (%) for encoder-only MRC systems subjecting to natural perturbations. In SQuAD 2.0, the values shown in the parentheses represent the relative change for answerable and unanswerable questions, respectively.

5.2 Error Analysis

Although encoder-only MRC models exhibit a relatively small performance gap, it remains worthwhile to investigate the sources of natural perturbation and reveal the perturbation phenomena contributing to models' error. To this end, within the original and the perturbed test set pair generated using the "Nature" method based on SQuAD 2.0 development set, we first identify 384 instances where at least one encoder-only model succeeds on the original but fails[5] on the perturbed (i.e., being adversarial), and then randomly select the same num-

[4]We identify a collection of such phrases by manually examining the decoder-only models' outputs (Check Appendix G for the full set).

[5]For answerable questions, a model's prediction is considered correct if both the EM and F1 scores are 1, and incorrect if both metrics are 0 or it determines the question is unanswerable. For unanswerable questions, a model's prediction is correct if it predicts the question is unanswerable, and wrong if it provides an answer span.

ber of instances on which all encoder-only models succeed on both the original and perturbed versions (Naik et al., 2018). We refer to these two types of instances as C2W (correct to wrong) and C2C (correct to correct) instances, respectively. Among the identified C2W and C2C instances, we further remove duplicates, resulting in 210 and 244 unique original and perturbed paragraph pairs, respectively. Furthermore, as natural perturbation can occasionally help the model to get the answer correct, we also filter 85 unique W2C (wrong to correct) instances on which at least two encoder-only models fail on the original but succeed on the perturbed. Finally, utilising an 8-category taxonomy of the semantic edit intentions in Wikipedia revisions derived from Yang et al. (2017), the first author of the paper manually annotated the chosen 210 samples of C2W and C2C, as well as the 85 W2C samples. To validate our findings, we further present 20% of the annotated C2W and C2C examples to a second annotator. See Appendix H for the instruction provided to the annotators, along with detailed explanations of each edit intention. We calculate the (micro-averaged) F1 score to evaluate the inter-annotator agreement, which is 0.82. This suggests that the annotators' annotations align closely. Table 3 reports the annotation results.

Edit Intention	C2W	C2C	W2C
Copy Editing	43.3	47.1	40.0
Clarification	5.7	3.3	1.2
Elaboration	23.8	18.1	22.4
Fact Update	4.3	3.8	3.5
Refactoring	1.9	1.9	1.2
Simplification	14.3	8.6	21.2
Vandalism	21.0	17.6	23.5
Other	9.5	14.8	10.6

Table 3: The percentage (%) of samples annotated with each edit intention in the C2W, C2C and W2C categories. The percentages do not add up to 100% because a single revision may fall into multiple intentions.

From Table 3, we observe that there is no significant difference in the distribution of annotated edit intentions between C2W and C2C examples, suggesting that ***though these types of natural perturbations confuse the encoder-only MRC models, the effect is not as pronounced***. A roughly similar distribution is also observed in the W2C examples, which indicates that these natural perturbation types can also facilitate correct answers by the models, i.e., being beneficial. Copy editing constitutes the most frequent edit intention (more than 40%), followed by elaboration and vandalism, with refactoring represents the category with the lowest percentage. Moreover, we find that there might be no correlation between the quality of the perturbed passage and its potential for being adversarial in the MRC robustness evaluation. Certain text edits aimed at improving the passage quality, such as copy editing and elaboration, do render the perturbation adversarial, whereas edits intended to damage the article may not consistently result in adversarial instances; in fact, vandalism can even assist models in providing correct answers. Instead, we infer that whether an edit to the passage can render the MRC instance adversarial or not depends on the location of the edits in relation to the question. Among the 384 C2W and C2C examples, we measure the proportion of answerable questions with the answer sentence(s) in the original passage remaining unmodified in the naturally perturbed version, which is 34.5% and 71.5%, respectively. This confirms our hypothesis that if the edits affect the answer sentence(s), there is a higher likelihood of the perturbed example becoming adversarial; otherwise, it might not. Appendix I presents one perturbed example for each of the C2W, C2C, and W2C categories, respectively, along with the annotated natural perturbation type(s).

5.3 Validity of Nature Adversarial Examples

To accurately assess a model's robustness under perturbation, it is vital to examine the validity of adversarial example, i.e. whether humans can still find the correct answer under the perturbation (Dyrmishi et al., 2023). We first present two human annotators with the same collection of adversarial instances, which includes only perturbed contexts and their corresponding questions, and then ask them to answer the question based on the perturbed context. The annotators are required to select the shortest continuous span in the perturbed context that answers the question and are allowed to leave the answer blank if they are confident that the question is not answerable. Full instructions given to the annotators can be seen in Appendix H. Subsequently, for both annotators, we measure the correctness (1 or 0) of their provided answers by comparing each of them with the corresponding ground truth answers[6]. The inter-annotator agree-

[6]Here, as long as one of the ground truth answers is included in the human-provided answer span, we consider the

ment is then measured by computing the Cohen's κ coefficient (Cohen, 1960). We then involve a third human annotator to annotate the adversarial examples on which the first two annotators disagree and then take the majority label as ground truth.

We employ this approach to verify the validity of the 210 C2W examples in Section 5.2 and find that 86% of these adversarial examples are valid (0.77 Cohen's κ), indicating that ***a substantial proportion of natural adversarial examples for encoder-only MRC model(s) are valid***.

5.4 Can Errors from Encoder-only Models Affect Other Architectures?

We are also curious about how well the errors identified in encoder-only models carry over to other model architectures. This leads us to first propose an exhaustive search algorithm that leverages the predictions of all encoder-only models to create the challenging natural perturbed test set. In detailed terms, for each matched reading passage from the prior version and its counterpart from the current version, we determine which should be designated as the *original* and which as the *perturbed* based on which scenario can yield the questions on which the maximum sum of the number of encoder-only models demonstrates the lack of robustness phenomenon[7]. Questions on which none of the encoder-only models fail under the perturbation are then removed. We finally process the identified original and perturbed passage pairs to ensure that the original passages are within the original SQuAD 1.1 development set. For those original passages with multiple occurrences, we select the one with the maximum number of questions reserved. With the development set of SQuAD 1.1 and SQuAD 2.0 as the source, this results in two challenge perturbed test sets: NAT_V1_CHALLENGE and NAT_V2_CHALLENGE. In NAT_V1_CHALLENGE, there are 184 contexts and 234 questions. NAT_V2_CHALLENGE contains 214 contexts and 442 questions (226 unanswerable).

Table 4 shows the evaluation results of both encoder-decoder and decoder-only models on the newly generated challenge test sets. From the table, we observe that ***the errors caused by natural perturbation in encoder-only MRC models transfer to both Flan-T5 and LLMs***. On the NAT_V1_CHALLENGE, `Flan-T5-small` demonstrates the greatest susceptibility to natural perturbation, experiencing a 14.27% decrease in F1, while among LLMs, `Gemma-7B-IT` emerges as the least robust, with a 16.66% F1 drop. Transitioning to the NAT_V2_CHALLENGE, the base version of `Flan-T5` exhibits the largest performance decline (13.83%) and `Falcon-7B-Instruct` stands out as the LLM with the lowest robustness. In Appendix J, we showcase two adversarial examples targeting LLMs sourced from our generated challenge sets.

Model	Performance *original* vs. *perturbed*			
	NAT_V1_CHALLENGE		NAT_V2_CHALLENGE	
`flan-t5-small`	58.76/64.76	48.58/55.52$_{-14.27}$	42.57/44.57	39.71/41.81$_{-6.19}$
`flan-t5-base`	79.49/85.01	66.1/73.42$_{-13.63}$	70.66/72.85	61.16/62.78$_{-13.83}$
`flan-t5-large`	88.1/92.53	76.57/82.31$_{-11.05}$	79.11/81.01	70.14/72.13$_{-10.96}$
`flan-t5-xl`	86.25/91.57	75.0/81.45$_{-11.05}$	83.71/85.84	73.19/74.86$_{-12.79}$
`GPT-3.5-turbo-0125`	91.03	83.33$_{-8.46}$	51.58	47.06$_{-8.76}$
`Gemma-2B-IT`	51.28	43.16$_{-15.83}$	55.66	50.23$_{-9.76}$
`Gemma-7B-IT`	82.05	68.38$_{-16.66}$	59.95	57.01$_{-4.9}$
`Llama 2-chat-7B`	82.91	73.93$_{-10.83}$	41.63	38.69$_{-7.06}$
`Llama 2-chat-13B`	80.77	73.93$_{-8.47}$	46.83	41.18$_{-12.06}$
`Llama-3-8B-Instruct`	88.89	77.35$_{-12.98}$	51.81	46.61$_{-10.04}$
`Mistral-7B-Instruct-v0.2`	85.9	76.92$_{-10.45}$	55.43	52.04$_{-6.12}$
`Falcon-7B-Instruct`	53.42	50$_{-6.4}$	32.81	23.53$_{-28.28}$
`Falcon-40B-Instruct`	69.66	62.82$_{-9.82}$	38.69	36.88$_{-4.68}$

Table 4: The performance (%) of encoder-decoder and decoder-only MRC models on the newly generated original and naturally perturbed challenge test sets. Values in smaller font are changes (%) relative to the original performance of the model.

6 MRC Under Synthetic Perturbation

In Appendix K, we present the evaluation results of different levels of synthetic perturbations against all MRC model architectures on the correspondingly generated test dataset. It can be seen from this table that generally, ***MRC systems exhibit a lack of robustness to synthetic perturbations, with varying degrees of performance decline***. Methods AddSent, WSplit and WInsert (WE) lead to noticeable drops in model performance, whereas other techniques demonstrate relatively limited impact.

For each method, from its created SQuAD 2.0-format test set pair, we also randomly select 50 instances where the `GPT-3.5-turbo-0125` shows evidence of being not robust, resulting in a total of 800 adversarial examples. We then measure their validity using the methodology described in Section 5.3, shuffling their order to mitigate potential bias, and present the results (0.81 Cohen's κ) in Table 5. From Table 5, we can see that in general, character and word-level perturbation methods result in more valid adversarial

prediction to be correct.

[7]A model lacks robustness to the perturbation if it achieves 1 EM on the original but attains less than 0.4 F1 on the perturbed (for answerable questions).

examples than sentence and document-level approaches, even though certain methods are lacking validity, such as WSubstitute (CWE) (48%) and CharSwapRand (52%). WSplit achieves the highest attack validity with 74%, while AddSent attains the lowest with 28%, despite causing the largest performance decrease for the `GPT-3.5-turbo-0125`. This suggests that the AddSent method, while impactful, might frequently generate perturbed MRC instances on which even humans find challenging. In Appendix L, we demonstrate some valid synthetic adversarial examples.

Attack	Answered Correctly
CharOCR	64
CharInsert	70
CharSubstitute	56
CharSwapMid	60
CharSwapRand	52
WInsert (CWE)	64
WSubstitute (CWE)	48
WSplit	74
WSwap	60
WDelete	60
WCrop	68
WSynSub	62
WInsert (WE)	58
AddSent	28
DocPara	48
Style Transfer	46

Table 5: The percentage (%) of adversarial MRC instances correctly labelled by humans for each synthetic perturbation method.

7 Adversarial Training

To enhance model robustness, we conduct adversarial training by identifying six encoder-only model architectures that exhibit the highest robustness to natural perturbations in their respective categories (except `albert-xxlarge-v2` on NAT_V2_CHALLENGE), and presenting them with both original training data and the generated naturally perturbed training examples. We extract the entire Wikipedia revision histories for the 392 articles in the original SQuAD training set, and then obtain 5,262 (with 22,033 questions) and 5,311 (with 32,993 questions) perturbed contexts to augment the original SQuAD 1.1 and SQuAD 2.0 training set, respectively, using the methodology described in Section 3.1. Table 6 compares the performance of these models on NAT_V1_CHALLENGE and NAT_V2_CHALLENGE, before and after retraining. Further, we also evaluate the behavior of the retrained models on the constructed synthetically perturbed test sets and quantify the discrepancy from the performance achieved prior to retraining. The results are shown in Appendix M.

Model	Performance *original* vs. *perturbed*			
	NAT_V1_CHALLENGE		NAT_V2_CHALLENGE	
`distilbert-base`	64.53/70.45	41.03/47.6$_{-32.43}$	56.56/59.08	41.18/43.3$_{-26.71}$
	57.26/63.44	43.59/51.87$_{-18.24}$	53.17/55.4	43.89/45.51$_{-17.85}$
`bert-large-cased`	79.06/83.66	63.68/70.23$_{-16.05}$	66.29/68.35	53.17/55.04$_{-19.47}$
	74.79/80.14	59.83/67.5$_{-15.77}$	67.87/69.31	58.37/59.53$_{-14.11}$
`spanbert-large-cased`	84.19/88.2	67.95/74.77$_{-15.23}$	78.73/80.68	62.44/64.99$_{-19.45}$
	82.48/86.6	69.66/76.05$_{-12.18}$	78.28/80.0	65.61/67.12$_{-16.1}$
`roberta-large`	86.75/90.21	73.93/79.47$_{-11.91}$	82.13/84.27	66.29/68.52$_{-18.69}$
	83.33/87.15	70.94/76.53$_{-12.19}$	81.22/82.67	70.59/71.84$_{-13.1}$
`albert-xxlarge-v2`	84.62/89.64	73.93/78.77$_{-12.13}$	84.62/86.07	68.1/69.61$_{-19.12}$
	86.32/90.93	75.64/81.07$_{-10.84}$	82.58/84.08	70.59/72.78$_{-13.44}$
`deberta-large`	88.46/92.5	73.5/78.48$_{-15.16}$	85.07/86.65	71.49/73.0$_{-15.75}$
	88.03/91.84	76.92/81.53$_{-11.23}$	83.03/85.1	72.62/74.48$_{-12.48}$

Table 6: Comparison of the performance of several encoder-only MRC systems on NAT_V1_CHALLENGE and NAT_V2_CHALLENGE, before and after retraining. The results shown in the shaded areas represent the performance of the model retrained on the augmented training set with naturally perturbed instances.

Overall, we observe that retraining enhances both the performance of the models on the naturally perturbed test set and their robustness to natural perturbations, albeit causing a slight decrease in the original performance. However, the phenomenon of improved perturbed performance does not generally apply to most synthetic perturbations, which indicates that natural and synthetic perturbations might indeed be different.

8 Conclusion

In this paper, we mainly study the robustness of MRC models to natural perturbations, which occur under real-world conditions without intentional human intervention. Using the proposed evaluation framework, we show that certain naturally perturbed examples can indeed be adversarial, i.e., lead to model failure, even when the modifications aim to improve the overall passage quality. Natural perturbations also appear to differ significantly from synthetic ones, exhibiting a wide range of rich linguistic phenomena and may be more effective in generating valid adversarial instances. Adversarial training via augmentation with naturally perturbed samples is generally beneficial for enhancing the model's robustness to natural perturbations; yet, it is not particularly successful in handling most synthetic noises. Future work includes the exploration of alternative natural perturbation approaches and the design of more effective defensive strategies against both natural and synthetic attacks.

Limitations

We acknowledge the presence of several limitations in our work: *(i)* There is a need to expand our study to other MRC datasets to make the findings more generalisable; *(ii)* Our natural perturbation framework only works with Wikipedia-based benchmarks. Therefore, it is necessary to develop other methods that can introduce real-world textual perturbations; *(iii)* It is essential to design better prompts to enhance the alignment of certain LLM outputs with given instructions (particularly for unanswerable questions), thereby ensuring more accurate evaluation results. There is also a necessity to examine the robustness of LLMs using different prompting strategies such as few-shot in-context learning; *(iv)* Since the impact of training data augmentation is relatively limited, we need to explore better techniques to improve the robustness of encoder-only models to natural perturbations and further investigate the robustness connection between natural and synthetic perturbations. Enhancing the robustness of LLMs is also a potential future direction.

References

Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. 2022. Exploring robustness connection between artificial and natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 179–186.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yu Cao, Dianqi Li, Meng Fang, Tianyi Zhou, Jun Gao, Yibing Zhan, and Dacheng Tao. 2022. TASA: Deceiving question answering models by twin answer sentences attack. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11975–11992, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022a. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

Yangyi Chen, Hongcheng Gao, Ganqu Cui, Fanchao Qi, Longtao Huang, Zhiyuan Liu, and Maosong Sun. 2022b. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11222–11237, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Salijona Dyrmishi, Salah Ghamizi, and Maxime Cordy. 2023. How do humans perceive adversarial text? a reality check on the validity and naturalness of word-based adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8822–8836, Toronto, Canada. Association for Computational Linguistics.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271.

Xanh Ho, Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2023. A survey on measuring and mitigating reasoning shortcuts in machine reading comprehension. *Preprint*, arXiv:2209.01824.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 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.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe, Frédéric Blain, Constantin Orăsan, and Lucia Specia. 2021. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 625–638, Online. Association for Computational Linguistics.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 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)*, pages 1109–1121, Online. Association for Computational Linguistics.

Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. 2021. Why machine reading comprehension models learn shortcuts? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 989–1002, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Mosh Levy, Shauli Ravfogel, and Yoav Goldberg. 2023. Guiding LLM to fool itself: Automatically manipulating machine reading comprehension shortcut triggers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8495–8505, Singapore. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Riviere, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl'ement Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran, Minh Giang, Cl'ement Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael

Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Anibal Pedraza, Oscar Deniz, and Gloria Bueno. 2022. Really natural adversarial examples. *International Journal of Machine Learning and Cybernetics*, 13(4):1065–1077.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 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)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2021. Semantics altering modifications for evaluating comprehension in machine reading. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13762–13770.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2023. A survey of methods for revealing and overcoming weaknesses of data-driven natural language understanding. *Natural Language Engineering*, 29(1):1–31.

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2021. Benchmarking robustness of machine reading comprehension models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 634–644, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Son Quoc Tran, Phong Nguyen-Thuan Do, Uyen Le, and Matt Kretchmar. 2023. The impacts of unanswerable questions on the robustness of machine reading comprehension models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1543–1557, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiayi Wang, Rongzhou Bao, Zhuosheng Zhang, and Hai Zhao. 2022a. Distinguishing non-natural from natural adversarial samples for more robust pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 905–915, Dublin, Ireland. Association for Computational Linguistics.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022b. Measure and improve robustness in NLP models: A

survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Johannes Welbl, Pasquale Minervini, Max Bartolo, Pontus Stenetorp, and Sebastian Riedel. 2020. Undersensitivity in neural reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1152–1165, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Winston Wu, Dustin Arendt, and Svitlana Volkova. 2021. Evaluating neural model robustness for machine comprehension. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2470–2481, Online. Association for Computational Linguistics.

Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro. 2023a. Are machine reading comprehension systems robust to context paraphrasing? In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–196, Nusa Dua, Bali. Association for Computational Linguistics.

Yulong Wu, Viktor Schlegel, Daniel Beck, and Riza Batista-Navarro. 2023b. MMT's submission for the WMT 2023 quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 856–862, Singapore. Association for Computational Linguistics.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the robustness of reading comprehension models to entity renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520, Seattle, United States. Association for Computational Linguistics.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations*.

A Datasets Statistics

We demonstrate the number of titles, contexts, and questions contained in the studied MRC datasets and the constructed perturbed test sets in Table 7.

Method	SQuAD 1.1			SQuAD 2.0		
	titles	contexts	questions	titles	contexts	questions$_{(HAns/NAns)}$
training	442	18896	87599	442	19035	130319$_{(86821/43498)}$
development	48	2067	10570	35	1204	11873$_{(5928/5945)}$
Nature	44	674	2776	32	368	3174$_{(1508/1666)}$
CharOCR	48	2007	6724	35	1194	7983$_{(3808/4175)}$
CharInsert	48	1952	5928	35	1189	7137$_{(3355/3782)}$
CharSubstitute	48	1956	5925	35	1181	7114$_{(3371/3743)}$
CharSwapMid	48	1956	6053	35	1188	7257$_{(3423/3834)}$
CharSwapRand	48	1952	5937	35	1190	7206$_{(3416/3790)}$
WInsert (CWE)	48	2030	7622	35	1203	9037$_{(4297/4740)}$
WSubstitute (CWE)	48	2005	6943	35	1200	8234$_{(3915/4319)}$
WSplit	48	1352	2477	35	993	3121$_{(1374/1747)}$
WSwap	48	1987	6257	35	1189	7442$_{(3516/3926)}$
WDelete	48	1982	6218	35	1190	7462$_{(3534/3928)}$
WCrop	48	2038	7459	35	1200	8809$_{(4240/4569)}$
WSynSub	48	1791	4970	35	1136	5427$_{(2546/2881)}$
WInsert (WE)	48	1840	4792	35	1166	5816$_{(2644/3172)}$
AddSent	48	2067	2075	35	1204	1205$_{(596/609)}$
DocPara	48	1968	6759	35	1188	7677$_{(3615/4062)}$
Style Transfer	48	1968	6740	35	1196	7753$_{(3647/4106)}$

Table 7: Dataset statistics of the SQuAD 1.1 and SQuAD 2.0, along with the respective perturbed test sets generated based on the development set of each. HAns: answerable questions; NAns: unanswerable questions.

B Visualisation of the Average Percentage of Editing Patterns

Figure 1 shows the average percentage of editing patterns identified across the articles contained within the development set of the investigated MRC datasets.

C Prompts for Sentence-Level and Document-Level Perturbations

In this section, we provide the prompts developed for implementing synthetic perturbations at both the sentence-level and document-level.

AddSent: *Generate a sentence unrelated to the context that shares significant lexical overlap with the given question but is not an answer to it.*\n\n*Question: {question}*\n\n*Context: {context}*

DocPara: *Given the context, paraphrase it as much as possible while still preserving the original*

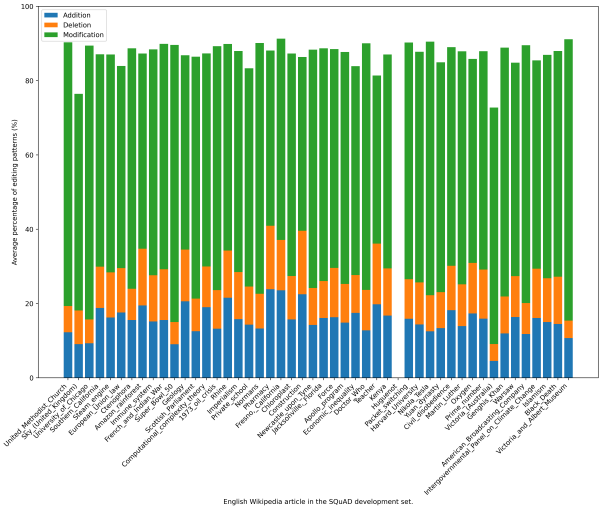

Figure 1: Average percentage of editing patterns for each Wikipedia article in the development set of our examined MRC datasets.

Model$_{Parameters(M)}$	d	b	lr	ep
DistilBERT$_{(66)}$	384	8	3e-5	3
BERT$_{(110/340)}$	384	8	3e-5	2
SpanBERT$_{(110/340)}$	512	4	2e-5	4
RoBERTa$_{(125/355)}$	384	8	3e-5	2
ALBERT$_{(11/17/223)}$	384	4	3e-5	2
DeBERTa$_{(350)}$	384	4	3e-6	3

Table 8: Number of parameters in each type of pre-trained encoder-only MRC model and the hyperparameters used to fine-tune them. For BERT, SpanBERT, RoBERTa and ALBERT, we show the number of model parameters in the order of *base*, *large* and *xxlarge* (if applicable) version. d is the size of the token sequence fed into the model, b is the training batch size, lr is the learning rate, and ep is the number of training epochs. We used stride = 128 for documents longer than d tokens.

meaning. Make sure to keep the phrases {ground truth answers[8]*} in the answer.*\n\n*Context: {context}*\n\n*The format of the output should be as follows:*\n\n*Rephrased Context:*

Style Transfer: *Given the context, identify a persona that can be used to rewrite it in a manner that results in a rephrased context that is both natural and realistic within a real-world scenario, while still preserving the original meaning. Make sure to keep the phrases {ground truth answers} in the answer.*\n\n*Context: {context}*\n\n*The format of the output should be as follows:*\n\n*Persona:*\n\n*Rephrased Context:*

D Encoder-only Model Parameters and Hyperparameters for Fine-tuning

Table 8 shows the hyperparameters used to fine-tune the pre-trained encoder-only MRC models in this work and their number of parameters contained.

E Instruction Templates for Flan-T5 Evaluation

In Table 9, we present the instruction templates employed in constructing the inputs to the Flan-T5 model for the SQuAD 1.1 format and SQuAD 2.0 format datasets, respectively.

[8]To be precise, here we are referring to all ground truth answers annotated for the questions that belong to the given context.

F MRC Prompts

We use the following two zero-shot prompts to instruct the decoder-only models to generate responses in the task of MRC, designed for the test sets with SQuAD 1.1 and SQuAD 2.0 format, respectively.

SQuAD 1.1: *Use the provided article delimited by triple quotes to answer question. Provide only the shortest continuous span from the context without any additional explanation.*\n\n*"""{context}"""*\n\n*Question: {question}*

SQuAD 2.0: *Use the provided article delimited by triple quotes to answer question. Provide only the shortest continuous span from the context without any additional explanation. If the question is unanswerable, return "unanswerable".*\n\n*"""{context}"""*\n\n*Question: {question}*

G Indicators of Unanswerable

We manually identify a set of phrases contained in the output of LLMs that indicate the unanswerability of the question, including *"I cannot answer this/the question", "unanswerable", "There is no indication in the provided article", "The context provided does not provide enough information", "There is no reference in the given article", "The answer to the question is not provided in the given article", "it is not possible", "question cannot be answered"* and *"context/question/article/text/article provided/passage does not"*.

	SQuAD 1.1
1	“Read this and answer the question\n\n{context}\n\n{question}”
2	“{context}\n{question}”
3	“Answer a question about this article:\n{context}\n{question}”
4	“Here is a question about this article: {context}\nWhat is the answer to this question: {question}”
5	“Article: {context}\n\nQuestion: {question}”
6	“Article: {context}\n\nNow answer this question: {question}”
	SQuAD 2.0
1	“Read this and answer the question. If the question is unanswerable, say \"unanswerable\".\n\n{context}\n\n{question}”
2	“{context}\n{question} (If the question is unanswerable, say \"unanswerable\")”
3	“{context}\nTry to answer this question if possible (otherwise reply \"unanswerable\"): {question}”
4	“{context}\nIf it is possible to answer this question, answer it for me (else, reply \"unanswerable\"): {question}”
5	“{context}\n\nAnswer this question, if possible (if impossible, reply \"unanswerable\"): {question}”
6	“Read this: {context}\nNow answer this question, if there is an answer (If it cannot be answered, return \"unanswerable\"): {question}”

Table 9: Various instruction templates for the `Flan-T5` model evaluation on the two benchmark MRC datasets.

H Human Annotation Instructions

In Figure 2, we show the instructions given to human annotators for error analysis (Section 5.2) and adversarial validity checking (Section 5.3), respectively. All our human annotators are students from universities in the United Kingdom and China. Before commencing each task, we ask them to annotate some examples and report the average time spent on each. As compensation, annotators receive 40 pence for each annotated example.

I Demonstration of Perturbed MRC Examples for Encoder-only Models

Figure 3 illustrates a naturally perturbed MRC instance each for categories C2W, C2C, and W2C, with the annotated perturbation type(s).

Error Analysis

You will be presented with pairs of reading contexts and their modified versions. The task is to compare each context and its modified version, observe the changes made and classify them into one or more of the semantic edit intention categories detailed below:

- *Copy Editing*: Rephrase; improve grammar, spelling, tone, or punctuation
- *Clarification*: Specify or explain an existing fact or meaning by example or discussion without adding new information
- *Elaboration*: Extend/add new content; insert a fact or new meaningful assertion
- *Fact Update*: Update numbers, dates, scores, episodes, status, etc. based on newly available information
- *Refactoring*: Restructure the article; move and rewrite content, without changing the meaning of it
- *Simplification*: Reduce the complexity or breadth of discussion; may remove information
- *Vandalism*: Deliberately attempt to damage the article
- *Other*: None of the above

We will use your annotation to calculate the percentage of each edit category.

Adversarial Validity Checking

Please read each provided context carefully and answer a corresponding question. Select the shortest continuous span from the context as your answer. If you believe a question cannot be answered, leave the answer blank. Your answer will be compared with the ground truth answers, and the result will only be used to decide the human answerability of the question.

Figure 2: Instructions for the two distinct human annotation tasks. In the error analysis task, the eight semantic edit intentions are adopted from (Yang et al., 2017).

Category: C2W

Original Paragraph: *Jacksonville, like most large cities in the United States, suffered from negative effects of rapid urban sprawl after World War II. The construction of highways led residents to move to newer housing in the suburbs. After World War II, the government of the city of Jacksonville began to increase spending to fund new public building projects in the boom that occurred after the war.* [. . .]
Perturbed Paragraph: *Jacksonville, like most large cities in the United States, suffered from negative effects of rapid urban sprawl after World War V. The construction of highways led residents to move to newer housing in the suburbs. After World War II, the government of the city of Jacksonville began to increase spending to fund new public building projects in the boom that occurred after the war.* [. . .]
Question: What did Jacksonville suffer from following World War I?
Prediction of `distilbert-base` and `spanbert-large-cased`: unanswerable→rapid urban sprawl
Annotated Natural Perturbation Type: Vandalism

Category: C2C

Original Paragraph: *Construction projects can suffer from preventable financial problems. Underbids happen when builders ask for too little money to complete the project. Cash flow problems exist when the present amount of funding cannot cover the current costs for labour and materials, and because they are a matter of having sufficient funds at a specific time, can arise even when the overall total is enough. Fraud is a problem in many fields, but is notoriously prevalent in the construction field. Financial planning for the project is intended to ensure that a solid plan with adequate safeguards and contingency plans are in place before the project is started and is required to ensure that the plan is properly executed over the life of the project.*
Perturbed Paragraph: *Financial planning ensures adequate safeguards and contingency plans are in place before the project is started, and ensures that the plan is properly executed over the life of the project. Construction projects can suffer from preventable financial problems. Underbids happen when builders ask for too little money to complete the project. Cash flow problems exist when the present amount of funding cannot cover the current costs for labour and materials; such problems may arise even when the overall budget is adequate, presenting a temporary issue. Fraud is also an occasional construction issue.*
Question: What can construction projects suffer from?
Prediction of all encoder-only models: preventable financial problems→preventable financial problems
Annotated Natural Perturbation Type: Copy Editing; Refactoring; Simplification

Category: W2C

Original Paragraph: *[. . .] The antigens expressed by tumors have several sources; some are derived from oncogenic viruses like human papillomavirus, which causes cervical cancer, while others are the organism's own proteins that occur at low levels in normal cells but reach high levels in tumor cells. [. . .] A third possible source of tumor antigens are proteins normally important for regulating cell growth and survival, that commonly mutate into cancer inducing molecules called oncogenes.*
Perturbed Paragraph: *[. . .] The antigens expressed by tumors have several sources; some are derived from oncogenic viruses like human papillomavirus, which causes cancer of the cervix, vulva, vagina, penis, anus, mouth, and throat,while others are the organism's own proteins that occur at low levels in normal cells but reach high levels in tumor cells. [. . .] A third possible source of tumor antigens are proteins normally important for regulating cell growth and survival, that commonly mutate into cancer inducing molecules called oncogenes.*
Question: What is a fourth possible source for tumor antigens?
Prediction of `bert-base-uncased`: proteins normally important for regulating cell growth and survival→unanswerable
Annotated Natural Perturbation Type: Elaboration

Figure 3: Natural perturbed MRC example in C2W, C2C and W2C categories.

J Natural Adversarial Samples for LLMs

We demonstrate two naturally perturbed reading comprehension examples that pose challenges for LLMs in Figure 4.

K Evaluation Results Under Synthetic Perturbation

The complete robustness evaluation results of the MRC models under synthetic perturbations are shown in Table 10.

L Synthetic Adversarial Instances

In Figure 5 and Figure 6, we present several synthetic adversarial samples that can be solved by humans.

M Robustness Connection Between Synthetic and Natural Perturbations

Figure 7 describes the impact of natural adversarial retraining on handling synthetic perturbations and will be included in the main body of the paper once we are given one additional page of content.

NAT_V1_CHALLENGE

Original Paragraph: *In business, notable alumni include Microsoft CEO Satya Nadella, Oracle Corporation founder and the third richest man in America Larry Ellison, Goldman Sachs and MF Global CEO as well as former Governor of New Jersey Jon Corzine, McKinsey & Company founder and author of the first management accounting textbook James O. McKinsey, Arley D. Cathey, Bloomberg L.P. CEO Daniel Doctoroff, Credit Suisse CEO Brady Dougan, Morningstar, Inc. founder and CEO Joe Mansueto, Chicago Cubs owner and chairman Thomas S. Ricketts, and NBA commissioner Adam Silver.*

Perturbed Paragraph: *In business, notable alumni include Microsoft CEO Satya Nadella, Oracle Corporation founder and the third richest man in America Larry Ellison, Goldman Sachs and MF Global CEO as well as former Governor of New Jersey Jon Corzine, McKinsey & Company founder and author of the first management accounting textbook James O. McKinsey, co-founder of the Blackstone Group Peter G. Peterson, co-founder of AQR Capital Management Cliff Asness, founder of Dimensional Fund Advisors David Booth, founder of The Carlyle Group David Rubenstein, Lazard CEO Ken Jacobs, entrepreneur David O. Sacks, CEO of TPG Group and former COO of Goldman Sachs Jon Winkelreid, former COO of Goldman Sachs Andrew Alper, billionaire investor and founder of Oaktree Capital Management Howard Marks, Bloomberg L.P. CEO Daniel Doctoroff, Credit Suisse CEO Brady Dougan, Morningstar, Inc. founder and CEO Joe Mansueto, Chicago Cubs owner and chairman Thomas S. Ricketts, and NBA commissioner Adam Silver.*

Question: What Goldman Sachs CEO is also an alumni of the University of Chicago?

Prediction of `GPT-3.5-turbo-0125` and `Llama-3-8B-Instruct`: Jon Corzine→Jon Winkelreid

Prediction of `Falcon-40B-Instruct`: Jon Corzine→David Rubenstein, co-founder of The Carlyle Group, is also an alumnus of the University of Chicago.

NAT_V2_CHALLENGE

Original Paragraph: *Each chapter has a number of authors who are responsible for writing and editing the material. A chapter typically has two "coordinating lead authors", ten to fifteen "lead authors", and a somewhat larger number of "contributing authors". The coordinating lead authors are responsible for assembling the contributions of the other authors, ensuring that they meet stylistic and formatting requirements, and reporting to the Working Group chairs. Lead authors are responsible for writing sections of chapters. Contributing authors prepare text, graphs or data for inclusion by the lead authors.*

Perturbed Paragraph: *Each chapter has a number of authors to write and edit the material. A typical chapter has two coordinating lead authors, ten to fifteen lead authors and a larger number of contributing authors. The coordinating lead authors assemble the contributions of the other authors. They ensure that contributions meet stylistic and formatting requirements. They report to the Working Group co-chairs. Lead authors write sections of chapters. They invite contributing authors to prepare text, graphs or data for inclusion.*

Question: Who has the responsibility for publishing materials?

Prediction of `Mistral-7B-Instruct-v0.2`: Unanswerable. The text does not mention any responsibility related to publishing materials.→The coordinating lead authors are responsible for publishing materials in the given context.

Figure 4: Natural perturbed MRC examples that confuse LLMs.

Dataset	Victim \ Attacker	CharOCR	CharInsert	CharSubstitute	CharSwapMid	CharSwapRand	WInsert (CWE)	WSubstitute (CWE)	WSplit	WSwap	WDelete	WCrop	WSynSub	WInsert (WE)	AddSent	DocPara	Style Transfer
SQuAD 1.1	distilbert-base	-5.72	-8.71	-8.27	-8.18	-8.28	-2.27	-3.59	-28.25	-2.05	-2.17	-1.33	-9.78	-25.93	-14.69	-6.79	-7.28
	bert-base-cased	-4.14	-6.6	-6.82	-6.04	-5.95	-2.22	-3.21	-21.57	-2.05	-2.29	-1.32	-8.67	-18.26	-22.95	-7.0	-6.59
	bert-base-uncased	-4.65	-7.85	-7.32	-7.3	-6.91	-2.18	-3.31	-26.02	-1.83	-2.2	-1.44	-9.25	-23.99	-24.08	-6.15	-6.63
	bert-large-cased	-2.04	-3.84	-4.01	-3.96	-4.2	-2.18	-2.94	-14.19	-1.38	-1.54	-1.32	-6.71	-12.67	-26.08	-6.7	-5.55
	bert-large-uncased	-2.78	-4.34	-5.11	-5.06	-4.53	-2.21	-3.34	-20.55	-1.16	-1.88	-1.42	-8.12	-18.55	-22.19	-5.82	-5.61
	spanbert-base-cased	-3.77	-6.48	-7.31	-7.06	-7.02	-2.45	-3.46	-24.71	-2.21	-2.34	-1.68	-8.84	-21.36	-17.3	-6.37	-5.73
	spanbert-large-cased	-2.7	-5.42	-5.66	-5.54	-5.48	-2.14	-3.19	-22.36	-1.65	-2.37	-1.51	-8.82	-18.32	-16.24	-6.41	-5.64
	roberta-base	-2.26	-3.77	-4.72	-4.19	-3.99	-1.88	-2.97	-12.19	-1.08	-1.84	-1.27	-6.11	-12.18	-13.98	-6.45	-5.42
	roberta-large	-1.57	-2.34	-2.5	-2.83	-3.08	-1.8	-2.83	-9.11	-1.09	-1.44	-0.93	-5.11	-9.01	-14.19	-5.65	-4.75
	albert-base-v1	-4.94	-7.54	-8.11	-7.59	-7.6	-2.52	-3.72	-27.79	-2.02	-2.61	-1.56	-8.65	-23.01	-25.92	-6.73	-6.92
	albert-base-v2	-3.79	-5.65	-6.47	-6.79	-6.24	-2.1	-3.46	-23.17	-1.87	-2.33	-1.59	-8.59	-18.75	-23.59	-7.19	-6.18
	albert-large-v1	-3.55	-5.22	-5.89	-6.27	-6.03	-2.08	-3.21	-20.56	-1.92	-2.23	-1.54	-8.75	-18.73	-19.52	-7.37	-6.47
	albert-large-v2	-3.33	-5.07	-5.31	-5.81	-6.05	-1.94	-3.1	-21.54	-1.86	-2.8	-1.39	-8.21	-18.76	-19.52	-6.96	-6.14
	albert-xxlarge-v1	-1.43	-2.06	-2.44	-2.82	-2.68	-1.7	-2.95	-10.4	-1.12	-1.84	-1.15	-4.94	-7.61	-15.31	-5.59	-4.68
	albert-xxlarge-v2	-1.28	-2.08	-2.44	-2.98	-2.77	-1.76	-2.75	-11.68	-1.11	-1.5	-1.27	-5.11	-9.03	-13.19	-5.24	-4.64
	deberta-large	-0.86	-1.54	-1.63	-1.78	-2.24	-1.36	-2.68	-7.23	-0.51	-1.09	-1.06	-3.95	-4.83	-11.52	-4.98	-4.45
	flan-t5-small	-5.31	-7.93	-8.83	-7.34	-7.49	-2.16	-5.02	-20.88	-2.95	-2.88	-1.49	-10.02	-19.66	-12.51	-7.85	-7.96
	flan-t5-base	-2.34	-3.32	-4.54	-4.36	-4.8	-2.06	-4.33	-11.95	-1.45	-2.7	-1.71	-8.29	-9.67	-11.54	-7.05	-6.03
	flan-t5-large	-1.16	-1.24	-2.01	-1.5	-2.15	-1.98	-4.79	-6.43	-0.72	-1.61	-1.29	-5.35	-4.08	-9.3	-5.97	-5.18
	flan-t5-xl	-1.14	-1.22	-1.94	-1.94	-2.23	-1.67	-4.44	-5.43	-1.12	-1.13	-1.06	-3.83	-1.92	-8.05	-5.35	-4.65
	GPT-3.5-turbo-0125	-0.14	-0.34	-0.59	-0.88	-0.69	-0.89	-1.74	0.17	-0.9	-1.0	-0.7	-4.14	-1.02	-9.22	-3.57	-3.82
	Gemma-2B-IT	-3.5	-5.46	-12.18	-6.53	-7.9	-3.62	-11.93	-12.47	-3.3	-8.88	-4.53	-39.87	-18.93	-4.57	-15.46	-14.99
	Gemma-7B-IT	-2.17	-2.59	-5.28	-3.96	-4.41	-1.38	-6.74	-3.69	-1.76	-5.36	-3.5	-14.97	-5.97	-11.43	-11.6	-10.25
	Llama 2-chat-7B	-0.82	-0.46	-1.36	-1.59	-1.89	-0.86	-3.16	-1.06	-0.83	-1.25	-0.64	-5.15	-2.94	-11.79	-4.5	-3.76
	Llama 2-chat-13B	-1.0	-0.7	-1.17	-1.35	-1.34	-0.61	-3.2	-0.51	-1.1	-1.54	-1.87	-4.61	-1.68	-6.97	-5.41	-4.21
	Llama-3-8B-Instruct	-0.49	-0.73	-1.52	-1.53	-1.43	-0.34	-3.18	-0.64	-0.99	-1.23	-1.1	-4.42	-1.43	-10.79	-5.5	-4.37
	Mistral-7B-Instruct-v0.2	-0.97	-0.42	-1.36	-0.96	-1.27	-0.31	-2.02	-1.46	-1.1	-1.77	-1.13	-4.75	-1.94	-5.25	-4.42	-3.83
	Falcon-7B-Instruct	-3.81	-6.25	-6.68	-6.43	-4.9	-2.73	-3.19	-9.37	-3.87	-3.61	-2.41	-12.98	-9.95	-8.01	-4.89	-2.75
	Falcon-40B-Instruct	-1.83	-0.73	-3.56	-2.37	-2.68	-1.77	-4.21	-0.88	-0.9	-2.61	-1.65	-6.91	-3.3	-14.66	-6.32	-4.32
SQuAD 2.0	distilbert-base	-2.65(−9.96/5.42)	-3.96(−19.14/12.57)	-5.38(−19.6/10.27)	-4.28(−18.33/10.99)	-5.16(−18.12/9.23)	-1.4(−4.41/1.98)	-2.63(−9.65/5.23)	-11.56(−59.55/35.71)	-0.64(−2.3/1.17)	-1.64(−6.0/3.18)	-1.36(−3.38/0.9)	-6.59(−23.67/12.97)	-10.56(−43.28/22.85)	-24.02(−21.13/−27.36)	-5.58(−16.31/6.58)	-4.28(−14.6/7.31)
	bert-base-cased	-3.73(−10.33/2.83)	-5.88(−17.58/5.53)	-5.47(−17.0/5.95)	-5.86(−16.27/4.36)	-5.84(−16.02/4.31)	-1.48(−3.68/0.74)	-3.88(−9.6/1.85)	-13.54(−47.25/15.11)	-0.91(−1.88/0.07)	-2.26(−6.11/1.54)	-1.52(−3.81/0.79)	-8.53(−21.69/4.83)	-9.72(−36.21/14.6)	-31.3(−37.28/−25.26)	-6.77(−16.31/2.85)	-6.9(−14.93/1.14)
	bert-base-uncased	-3.82(−11.45/3.91)	-6.02(−17.34/5.11)	-5.83(−17.29/5.64)	-5.27(−17.17/6.54)	-5.83(−16.85/5.26)	-1.03(−3.97/1.96)	-4.13(−11.19/3.03)	-15.7(−55.42/19.32)	-0.43(−1.59/0.73)	-2.4(−6.43/1.63)	-1.71(−4.43/1.05)	-9.42(−23.32/4.86)	-12.94(−36.69/9.24)	-33.02(−35.11/−30.86)	-6.45(−16.24/3.48)	-5.94(−14.64/2.78)
	bert-large-cased	-2.47(−6.61/1.65)	-4.86(−12.68/2.81)	-4.92(−12.64/2.78)	-5.52(−12.95/1.78)	-5.07(−12.15/2.01)	-1.2(−2.9/0.51)	-4.92(−11.17/1.35)	-15.1(−37.91/4.79)	-0.26(−1.15/0.61)	-3.21(−7.19/0.76)	-1.88(−4.28/0.56)	-10.34(−19.71/−0.91)	-6.77(−19.78/5.23)	-34.27(−46.81/−21.27)	-8.1(−15.52/−0.67)	-7.21(−13.01/−1.48)
	bert-large-uncased	-3.9(−8.21/0.36)	-5.34(−12.99/2.12)	-6.11(−13.99/1.69)	-5.61(−12.75/1.43)	-5.46(−12.48/1.5)	-1.66(−3.46/0.1)	-5.0(−11.07/1.02)	-16.73(−48.06/10.16)	-0.82(−2.08/0.4)	-2.75(−6.23/0.68)	-1.63(−3.72/0.48)	-10.98(−21.36/−0.51)	-11.27(−27.53/3.73)	-35.79(−44.61/−26.73)	-7.86(−15.76/0.0)	-7.54(−14.6/−0.57)
	spanbert-base-cased	-4.43(−8.05/−0.89)	-7.51(−14.59/−0.75)	-6.83(−14.95/0.96)	-7.11(−14.32/−0.2)	-7.84(−15.11/−0.86)	-1.74(−3.47/−0.05)	-5.56(−12.17/0.83)	-21.04(−39.6/−5.38)	-2.34(−2.16/−2.5)	-3.93(−6.44/−1.55)	-2.3(−3.74/−0.9)	-11.49(−17.9/−5.3)	-14.7(−26.02/−4.66)	-28.62(−32.9/−24.3)	-8.7(−15.36/−2.27)	-7.8(−15.05/−0.88)
	spanbert-large-cased	-3.91(−7.29/−0.74)	-7.08(−12.91/−1.67)	-6.83(−13.09/−1.0)	-7.32(−12.82/−2.21)	-7.46(−14.24/−1.11)	-1.49(−2.93/−0.17)	-5.7(−11.39/−0.35)	-18.52(−33.89/−6.04)	-2.14(−2.43/−1.86)	-3.4(−5.48/−1.46)	-2.02(−3.39/−0.71)	-11.39(−16.92/−6.24)	-13.11(−24.15/−3.59)	-29.65(−38.44/−21.03)	-8.6(−15.2/−2.46)	-7.42(−12.67/−2.55)
SQuAD 2.0	roberta-base	-3.05(−5.9/−0.29)	-5.45(−11.83/0.55)	-5.37(−12.15/1.11)	-5.52(−11.85/0.44)	-5.53(−12.68/1.29)	-1.22(−3.1/0.58)	-4.97(−11.87/1.64)	-15.02(−30.86/−1.85)	-0.99(−1.81/−0.21)	-3.56(−7.25/−0.06)	-2.15(−3.88/−0.46)	-8.84(−15.77/−2.2)	-9.32(0.23)	-25.73(−32.6/−18.85)	-8.05(−15.67/−0.78)	-7.69(−13.92/−1.8)
	roberta-large	-2.54(−4.11/−1.09)	-4.52(−7.69/−1.64)	-4.92(−8.19/−1.9)	-4.92(−7.76/−2.32)	-4.91(−8.44/−1.65)	-1.71(−3.41/−0.12)	-6.04(−12.93/0.36)	-11.91(−16.9/−7.92)	-1.55(−2.04/−1.1)	-3.5(−5.72/−1.44)	-2.28(−3.97/−0.67)	-10.58(−15.72/−5.83)	-8.69(−14.3/−3.9)	-25.66(−32.82/−18.52)	-8.33(−15.61/−1.63)	-7.37(−13.61/−1.63)
	albert-base-v1	-5.62(−12.18/0.23)	-8.7(−20.33/1.45)	-7.33(−19.14/3.14)	-8.17(−19.48/1.75)	-7.76(−18.59/1.98)	-1.76(−4.55/0.68)	-5.64(−12.99/0.9)	-20.95(−58.64/8.05)	-1.72(−4.08/0.34)	-4.62(−10.4/0.5)	-2.68(−5.56/−0.07)	-11.72(−23.83/−0.73)	-14.37(−33.25/1.12)	-36.85(−46.13/−28.13)	-9.64(−20.38/−0.09)	-8.71(−18.82/0.31)
	albert-base-v2	-5.17(−11.0/0.17)	-6.82(−16.17/1.51)	-7.91(−17.32/0.62)	-7.71(−16.9/0.55)	-7.14(−16.01/0.91)	-1.58(−4.05/0.65)	-5.84(−13.27/0.9)	-19.59(−52.41/6.15)	-2.0(−3.14/−0.99)	-4.09(−8.51/−0.08)	-2.61(−5.07/−0.38)	-10.99(−21.88/−0.99)	-11.7(−25.83/0.23)	-34.77(−41.78/−28.19)	-8.87(−17.83/−0.7)	-7.82(−16.47/0.02)
	albert-large-v1	-4.46(−10.28/0.93)	-7.91(−17.6/0.81)	-7.93(−18.27/1.46)	-7.69(−17.17/1.0)	-7.56(−17.03/1.13)	-1.67(−3.65/0.15)	-5.66(−13.24/1.36)	-19.28(−50.7/5.73)	-1.08(−3.08/0.75)	-3.75(−8.51/0.61)	-2.53(−5.72/0.43)	-12.02(−21.5/−3.11)	-13.36(−30.25/1.0)	-34.89(−40.81/−29.09)	-8.93(−17.81/−0.64)	-7.73(−16.31/0.21)
	albert-large-v2	-3.85(−8.34/0.38)	-6.45(−13.89/0.38)	-6.24(−13.93/1.01)	-6.34(−13.46/0.26)	-6.03(−12.75/0.26)	-1.46(−3.33/0.3)	-5.56(−12.35/0.88)	-19.77(−44.92/0.62)	-1.48(−2.72/−0.34)	-3.9(−8.15/0.06)	-2.69(−5.04/−0.45)	-11.43(−21.76/−1.6)	-10.57(−22.42/−0.27)	-31.77(−38.15/−25.49)	-8.91(−16.42/−1.88)	-8.2(−15.52/−1.37)
	albert-xxlarge-v1	-2.43(−2.92/−1.94)	-4.19(−5.59/−2.9)	-4.31(−5.05/−3.6)	-4.18(−5.77/−2.67)	-4.51(−7.14/−2.03)	-1.49(−2.64/−0.39)	-5.99(−11.76/−0.53)	-13.22(−18.41/−8.95)	-1.47(−1.62/−1.33)	-3.55(−5.68/−1.54)	-2.3(−4.02/−0.65)	-9.04(−12.65/−5.66)	-7.13(−12.51/−2.44)	-26.12(−28.58/−23.67)	-8.02(−14.92/−1.54)	-7.13(−12.9/−1.77)
	albert-xxlarge-v2	-2.25(−2.61/−1.9)	-3.91(−5.29/−2.59)	-4.51(−6.0/−3.09)	-5.05(−6.89/−3.32)	-4.86(−7.59/−2.25)	-1.46(−2.64/−0.35)	-5.68(−11.2/−0.39)	-13.52(−21.09/−7.21)	-1.3(−1.72/−0.91)	-3.23(−5.13/−1.44)	-1.96(−3.58/−0.4)	-8.8(−11.95/−5.82)	-7.33(−11.83/−3.37)	-25.06(−29.04/−21.07)	-8.41(−14.41/−2.73)	-6.92(−12.18/−1.98)
	deberta-large	-1.39(−1.62/−1.17)	-2.94(−2.37/−3.46)	-3.05(−3.5/−2.63)	-2.89(−3.19/−2.64)	-3.24(−4.18/−2.37)	-1.32(−2.6/−0.12)	-5.93(−12.1/−0.18)	-9.31(−10.38/−8.44)	-1.52(−1.6/−1.46)	-2.87(−4.39/−1.48)	-1.89(−3.37/−0.47)	-8.34(−9.08/−7.66)	-5.95(−5.89/−6.02)	-22.96(−27.86/−18.19)	-7.35(−13.84/−1.32)	-7.43(−12.99/−2.29)
	flan-t5-small	-2.07(−7.91/5.18)	-2.52(−10.98/7.92)	-2.37(−12.21/9.93)	-3.09(−11.09/6.77)	-2.28(−10.17/7.54)	-0.73(−3.73/3.05)	-2.05(−7.75/5.24)	-8.7(−31.68/15.87)	-0.12(−3.07/3.58)	-0.71(−5.68/5.5)	-1.27(−4.19/2.38)	-3.73(−24.18/23.62)	-7.04(−29.75/19.74)	-15.79(−15.74/−15.82)	-3.08(−17.91/16.18)	-2.88(−14.63/12.24)
	flan-t5-base	-2.35(−5.22/0.77)	-4.75(−9.59/0.36)	-4.85(−10.24/0.95)	-4.22(−9.03/0.9)	-4.77(−10.68/1.59)	-1.35(−3.27/0.73)	-4.34(−11.01/2.96)	-12.36(−21.26/−3.86)	-0.79(−1.65/0.11)	-3.22(−6.73/0.56)	-1.94(−4.19/0.51)	-10.13(−17.74/−1.79)	-8.01(−15.53/−0.4)	-25.97(−22.91/−29.41)	-7.63(−14.9/0.28)	-6.98(−11.69/−1.89)
	flan-t5-large	-2.23(−2.34/−2.15)	-4.09(−3.84/−4.32)	-4.07(−5.04/−3.15)	-3.69(−4.19/−3.21)	-4.24(−5.35/−3.18)	-1.7(−3.05/−0.42)	-5.95(−12.45/0.35)	-11.41(−8.14/−14.15)	-2.12(−1.03/−3.14)	-3.84(−5.83/−1.93)	-2.2(−3.16/−1.28)	-9.24(−11.81/−6.78)	-7.43(−5.96/−9.01)	-20.67(−20.33/−21.03)	-7.38(−14.85/−0.26)	-6.37(−12.06/−0.99)
	flan-t5-xl	-1.16(−1.68/−0.68)	-2.65(−3.48/−1.92)	-2.96(−4.18/−1.86)	-2.9(−4.26/−1.67)	-3.12(−4.99/−1.45)	-1.62(−3.32/−0.09)	-6.07(−12.88/0.08)	-6.12(−6.35/−5.95)	-1.49(−1.52/−1.48)	-3.06(−4.99/−1.32)	-2.33(−4.21/−0.61)	-7.06(−9.32/−5.03)	-3.19(−2.39/−3.87)	-19.91(−23.35/−16.64)	-7.54(−15.67/−0.3)	-7.08(−13.87/−1.07)
	GPT-3.5-turbo-0125	-1.31(−0.62/−4.66)	-3.6(−0.86/−15.8)	-2.81(−0.79/−12.15)	-3.34(−0.89/−14.7)	-3.53(−1.4/−13.76)	-1.66(−0.52/−6.92)	-1.57(−2.69/3.8)	-1.52(−1.07/−3.2)	-1.8(−0.9/−5.96)	-1.7(−1.06/−4.77)	-1.2(−0.78/−3.16)	-7.6(−5.63/−15.55)	-5.15(−1.51/−19.98)	-11.98(−13.27/−6.33)	-0.99(−4.99/16.57)	-1.52(−5.74/17.73)
	Gemma-2B-IT	-2.42(−4.44/−1.12)	-2.99(−9.18/0.78)	-3.44(−10.72/1.17)	-3.14(−7.72/−0.36)	-2.09(−6.26/0.56)	-1.72(−2.7/−1.1)	-2.82(−13.06/3.74)	-5.47(−20.84/2.2)	-1.31(−2.61/−0.52)	-1.59(−8.83/2.98)	-1.43(−4.59/0.7)	-8.54(−45.35/12.48)	-7.89(−19.54/−1.78)	-7.13(1.34/−13.83)	-1.29(−18.04/9.5)	-1.75(−16.17/7.45)
	Gemma-7B-IT	-2.64(−7.04/1.99)	-2.9(−9.48/3.81)	-3.88(−11.07/3.6)	-4.0(−9.92/2.06)	-3.6(−10.24/3.33)	-1.73(−3.33/−0.06)	-3.2(−9.16/2.95)	-2.2(−8.95/3.57)	-3.14(−6.29/0.03)	-1.7(−8.68/5.59)	-1.19(−4.77/2.7)	-8.15(−26.33/9.48)	-4.07(−12.84/3.95)	-12.09(−15.06/−8.66)	-2.79(−16.59/11.26)	-2.54(−15.18/10.3)
	Llama 2-chat-7B	-0.49(0.12/−8.16)	-2.75(−0.84/−25.26)	-2.86(−1.37/−21.38)	-2.34(−0.95/−19.33)	-2.72(−1.23/−21.9)	-0.94(−0.26/−8.97)	-3.56(−2.62/−15.04)	-3.26(−0.5/−28.52)	-0.96(−0.23/−10.13)	-1.68(−0.78/−12.77)	-1.07(−0.84/−3.86)	-6.49(−4.73/−28.48)	-6.11(−2.71/−44.26)	-8.72(−12.06/29.64)	-3.48(−5.16/16.82)	-3.2(−3.69/3.08)
	Llama 2-chat-13B	-2.76(−0.94/−9.44)	-3.32(−1.3/−10.15)	-4.63(−0.77/−17.78)	-3.58(−1.55/−10.68)	-4.3(−2.33/−11.23)	-1.02(−0.44/−3.11)	-3.83(−3.76/−4.08)	-2.23(−2.49/−1.47)	-1.82(−0.17/−7.52)	-3.53(−1.66/−10.08)	-1.41(−0.71/−4.02)	-10.46(−6.26/−23.67)	-5.81(−3.68/−12.88)	-6.92(−8.75/0.0)	-4.87(−5.11/−4.12)	-2.62(−4.05/2.29)
	Llama-3-8B-Instruct	0.05(−0.36/1.79)	-2.42(−1.46/−6.41)	-2.54(−1.82/−5.61)	-3.2(−2.27/−7.01)	-3.02(−2.21/−6.34)	-1.49(−0.57/−5.19)	-3.23(−3.57/−1.8)	-1.49(−1.31/−2.19)	-1.37(−0.69/−4.15)	-2.13(−1.85/−3.27)	-0.23(−0.9/2.54)	-10.32(−5.38/−28.89)	-8.19(−1.24/−33.72)	-15.5(−13.24/−25.18)	-1.62(−6.22/16.19)	-1.23(−5.3/14.4)
	Mistral-7B-Instruct-v0.2	-0.8(−0.97/−0.42)	-1.11(−1.21/−0.87)	-0.82(−1.08/−0.22)	-2.14(−1.54/−3.51)	-2.0(−2.16/−1.6)	-0.31(−0.49/0.14)	-2.27(−3.98/1.6)	-1.05(−1.25/−0.64)	-0.73(−0.5/−1.19)	-0.98(−2.28/1.88)	-0.28(−1.42/2.36)	-5.81(−6.1/−5.18)	-9.83(−1.56/−27.06)	-6.71(−10.44/1.69)	1.47(−6.43/19.04)	1.11(−4.88/14.35)
	Falcon-7B-Instruct	-1.6(−3.07/30.07)	-3.07(−3.99/3.32)	-5.59(−5.42/−6.66)	-5.95(−4.96/−13.24)	-5.05(−6.98/12.76)	1.03(0.09/8.61)	-5.41(−6.68/3.93)	-12.67(−13.56/−6.42)	-4.3(−4.32/−3.83)	-2.94(−3.42/0.74)	-2.01(−3.15/7.15)	-9.88(−12.32/13.25)	-7.89(−9.23/2.4)	-8.59(−11.81/11.38)	-5.31(−5.86/−1.36)	-0.29(−0.89/4.19)
	Falcon-40B-Instruct	-2.17(−2.26/−1.69)	-2.65(−2.33/−4.42)	-3.07(−2.62/−5.97)	-3.64(−3.09/−6.85)	-3.62(−2.98/−7.39)	-0.02(−0.28/1.39)	-4.06(−4.34/−2.37)	-1.47(0.0/−8.5)	-1.25(−1.07/−2.26)	-4.86(−6.4/4.84)	-1.66(−1.65/−1.78)	-8.54(−7.22/−16.2)	-5.53(−3.49/−16.99)	-10.53(−14.95/19.65)	-3.89(−6.03/8.24)	-5.28(−5.54/−3.9)

Table 10: Robustness of MRC models under synthetic perturbations. For encoder-only and encoder-decoder models, the results displayed are the relative F1 change (%), while for decoder-only models, we demonstrate the relative IM change (%). In SQuAD 2.0, the values shown in the parentheses represent the relative change for answerable and unanswerable questions, respectively.

CharSwapRand

Original Paragraph: *[. . .] According to the general principle of proportionality the lawfulness of an action depends on whether it was appropriate and necessary to achieve the objectives legitimately pursued. When there is a choice between several appropriate measures the least onerous must be adopted, and any disadvantage caused must not be disproportionate to the aims pursued. The principle of proportionality is also recognised in Article 5 of the EC Treaty, stating that "any action by the Community shall not go beyond what is necessary to achieve the objectives of this Treaty".*
Perturbed Paragraph: *[. . .] According to the general principle of proportionality the lawfulness of an action depends on whether it was appaoptirre and necessary to achieve the ibjvotcees legitimately psdeuur. When there is a choice between several ippraptoa seumraes the least onerous must be adopted, and any disadvantage caused must not be disproportionate to the aims upsured. The pricnplie of proportionality is also recognised in Article 5 of the EC Treaty, stating that " any acitno by the Community shall not go beyond what is necessary to achieve the objectives of this Treaty ".*
Question: Where is the principle of proportionality not recognized in the EC treaty?
Prediction of `GPT-3.5-turbo-0125`: unanswerable→In Article 5 of the EC Treaty.
Prediction by humans under perturbation: unanswerable

WSubstitute (CWE)

Original Paragraph: *One of the most famous people born in Warsaw was Maria SkÇodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include WÇadysÇaw Szpilman and Fr©d©ric Chopin. Though Chopin was born in the village of ªelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.*
Perturbed Paragraph: *One of the most famous Russians born outside Warsaw was Maria SkÇodowska-Curie, who achieved international recognition for her research on radioactivity who was its last female recipient granted the Mercury Prize. Famous musicians were WÇadysÇaw Szpilman and Fr©d©ric Chopin. Though Chopin was born in the village of ªelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he only seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was married here in 1745.*
Question: Who was Fredderic Wola?
Prediction of `GPT-3.5-turbo-0125`: unanswerable→Frederic Chopin.
Prediction by humans under perturbation: unanswerable

WSplit

Original Paragraph: *Civil disobedience is usually defined as pertaining to a citizen's relation to the state and its laws, as distinguished from a constitutional impasse in which two public agencies, especially two equally sovereign branches of government, conflict. For instance, if the head of government of a country were to refuse to enforce a decision of that country's highest court, it would not be civil disobedience, since the head of government would be acting in her or his capacity as public official rather than private citizen.*
Perturbed Paragraph: *C ivil disobedience is usually de fined as pertai ning to a citizen ' s relation to the st ate and its la ws, as distinguished from a constitutional impasse in which two pub lic age ncies, especially two eq ually sove reign b ranches of government, conflict. For ins tance, if the head of government of a c ountry were to refuse to en force a decision of that cou ntry ' s highest court, it would not be civil disobedience, since the he ad of government wo uld be acting in her or his capaci ty as p ublic offic ial rather t han private citizen.*
Question: What group promotes using conflict to decide cases?
Prediction of `GPT-3.5-turbo-0125`: unanswerable→public agencies
Prediction by humans under perturbation: unanswerable

Figure 5: Valid synthetic adversarial examples (character and word levels).

AddSent

Original Paragraph: *[. . .] The Anglo-Norman language was eventually absorbed into the Anglo-Saxon language of their subjects (see Old English) and influenced it, helping (along with the Norse language of the earlier Anglo-Norse settlers and the Latin used by the church) in the development of Middle English. It in turn evolved into Modern English.*
Perturbed Passage: *What was the final outcome of the Anglo-Norman language's influence on English literature?[. . .] The Anglo-Norman language was eventually absorbed into the Anglo-Saxon language of their subjects (see Old English) and influenced it, helping (along with the Norse language of the earlier Anglo-Norse settlers and the Latin used by the church) in the development of Middle English. It in turn evolved into Modern English.*
Question: What was the Anglo-Norman language's final form?
Prediction of `GPT-3.5-turbo-0125`: Modern English→It was eventually absorbed into the Anglo-Saxon language of their subjects.
Prediction by humans under perturbation: Modern English

DocPara

Original Paragraph: *The area is also known for its early twentieth century homes, many of which have been restored in recent decades. The area includes many California Bungalow and American Craftsman style homes, Spanish Colonial Revival Style architecture, Mediterranean Revival Style architecture, Mission Revival Style architecture, and many Storybook houses designed by Fresno architects, Hilliard, Taylor & Wheeler. The residential architecture of the Tower District contrasts with the newer areas of tract homes urban sprawl in north and east areas of Fresno.*
Perturbed Paragraph: *In recent decades, many early twentieth century homes in the north and east areas of Fresno have been restored, including California Bungalow and American Craftsman style properties, as well as Storybook houses designed by Hilliard, Taylor & Wheeler. The unique architectural styles of the Tower District, such as Spanish Colonial Revival, Mediterranean Revival, and Mission Revival, stand in contrast to the newer tract homes and urban sprawl in the surrounding areas.*
Question: Are California Bungalows located in the north or east?
Prediction of `GPT-3.5-turbo-0125`: unanswerable→In the north and east areas of Fresno.
Prediction by humans under perturbation: unanswerable

Style Transfer

Original Paragraph: *Western musical instruments were introduced to enrich Chinese performing arts. From this period dates the conversion to Islam, by Muslims of Central Asia, of growing numbers of Chinese in the northwest and southwest. Nestorianism and Roman Catholicism also enjoyed a period of toleration. Buddhism (especially Tibetan Buddhism) flourished, although Taoism endured certain persecutions in favor of Buddhism from the Yuan government. Confucian governmental practices and examinations based on the Classics, which had fallen into disuse in north China during the period of disunity, were reinstated by the Yuan court, probably in the hope of maintaining order over Han society. Advances were realized in the fields of travel literature, cartography, geography, and scientific education.*
Perturbed Paragraph: *During the Yuan dynasty, Western musical instruments were introduced to enhance Chinese performing arts, while the influence of Islam from Central Asia led to the conversion of some Chinese individuals in the northwest and southwest regions. Nestorianism and Roman Catholicism were also tolerated during this period. Despite the flourishing of Buddhism, Taoism faced persecution by the Yuan government in favor of Buddhism. The revival of Confucian governmental practices and examinations based on the Classics aimed to maintain order in Han society. Additionally, significant progress was made in travel literature, cartography, geography, and scientific education during this time.*
Question: What fields of study were not advanced during the Yuan?
Prediction of `GPT-3.5-turbo-0125`: unanswerable→Taoism
Prediction by humans under perturbation: unanswerable

Figure 6: Valid synthetic adversarial examples (sentence and document levels).

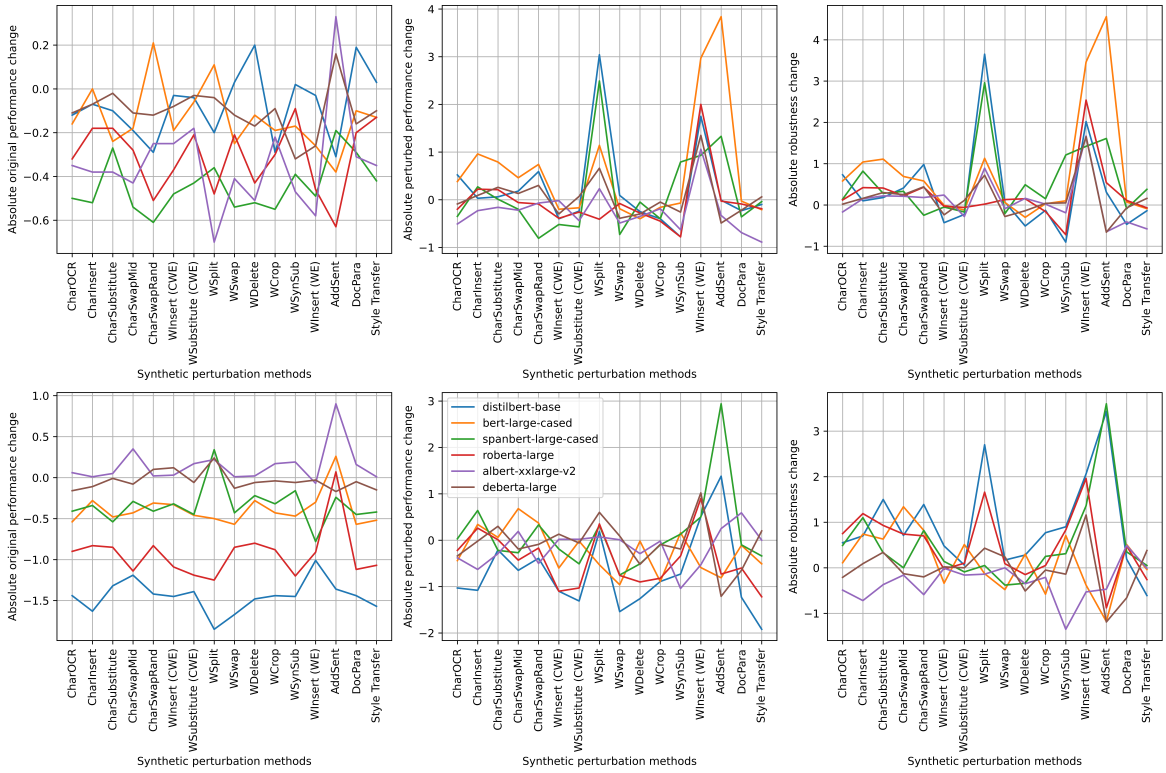

Figure 7: Absolute changes in original and perturbed performance (F1), as well as the robustness of six encoder-only models under various synthetic perturbations, following training on the augmented dataset with naturally perturbed MRC samples. The upper row of figures illustrates the outcomes obtained on the test sets created with SQuAD 1.1 as the reference, while the bottom row displays the results on SQuAD 2.0 format test datasets.