Are Machine Reading Comprehension Systems Robust to Context Paraphrasing?

Yulong Wu[1], **Viktor Schlegel**[1,2] **and Riza Batista-Navarro**[1]
[1] Department of Computer Science, University of Manchester, United Kingdom
[2] ASUS Intelligent Cloud Services (AICS), Singapore
{yulong.wu, riza.batista}@manchester.ac.uk
viktor_schlegel@asus.com

Abstract

Investigating the behaviour of Machine Reading Comprehension (MRC) models under various types of test-time perturbations can shed light on the enhancement of their robustness and generalisation capability, despite the superhuman performance they have achieved on existing benchmark datasets. In this paper, we study the robustness of contemporary MRC systems to context paraphrasing, i.e., whether these models are still able to correctly answer the questions once the reading passages have been paraphrased. To this end, we systematically design a pipeline to semi-automatically generate perturbed MRC instances which ultimately lead to the creation of a paraphrased test set. We conduct experiments on this dataset with six state-of-the-art neural MRC models and we find that even the minimum performance drop of all these models exceeds 41%, whereas human performance remains high. Retraining models with augmented perturbed examples results in improved robustness, though the performance remains lower than on the original dataset. These results demonstrate that the existing high-performing MRC systems are still far away from real language understanding[1].

1 Introduction

Machine reading comprehension (MRC), the task of automatically reading a passage of text and answering related questions, serves as an important testbed for evaluating various Natural Language Understanding (NLU) capabilities of computer systems (Chen, 2018). While neural MRC systems approach or even surpass human performance on benchmark datasets (Devlin et al., 2019; Lan et al., 2020; He et al., 2021), it remains uncertain whether they can indeed solve the MRC task (Schlegel et al., 2020; Wu et al., 2021b; Sugawara et al., 2022; Shinoda et al., 2023; Schlegel et al., 2023). In particular, recent studies have shown that instead of performing consistently well, contemporary models are brittle under various test-time perturbations (Ribeiro et al., 2020; Si et al., 2021; Wu et al., 2021a; Schlegel et al., 2021; Yan et al., 2022). This raises the question of the suitability of existing gold standard datasets to establish a model's robustness and the need to improve the reliability of these MRC systems (Wang et al., 2022).

Question: In what year did Harvard President Joseph Willard die?
Original Context: [...] When the Hollis Professor of Divinity David Tappan died in 1803 and the president of Harvard Joseph Willard died a year later, in 1804, a struggle broke out over their replacements. [...] **Prediction:** 1804
Paraphrased Context: [...] When the Hollis professor of divinity David Tappan died in 1803, the President of Harvard Joseph Willard died a year later, a battle broke out in 1804 for their successors. [...] **Prediction:** 1803 **Prediction by a human:** 1804

Figure 1: An instance where the RoBERTa-large model (Liu et al., 2019) can get the answer correct over the original reading passage, but is misled when presented with a whole context paraphrased version.

Paraphrase understanding plays a role in measuring the robustness and generalisation ability of MRC models. Intuitively, a trustworthy MRC system should demonstrate robust generalisation on paraphrased contexts and/or questions, i.e., those that convey the same semantic meaning using different surface forms. Previous studies have attempted to paraphrase the questions (Gan and Ng, 2019) and strategically modify portions of the reading passage, e.g., paraphrase only the answer sentence using the back-translation (Lai et al., 2021) or generate paraphrases that exclude the top five im-

[1]Our code and data are available at https://github.com/Yulong-W/context-paraphrasing.

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
November 1–4, 2023.

portant words in the context (Wu et al., 2021a). By assessing model performance on the paraphrased test sets, they concluded that MRC models might be vulnerable to paraphrasing-oriented attacks.

The reading comprehension task assesses a model's real understanding of a given context, i.e., a *passage*. Though the findings in the work of Lai et al. (2021) and Wu et al. (2021a) provide insights into the weaknesses of MRC datasets to benchmark partial-context paraphrasing understanding, their designed strategic paraphrasing approach may hinder the generated perturbed examples from accurately simulating real-world text disruptions, which can pervade any part of a passage, not just specific words or answer sentences. Furthermore, it is not clear whether the modifications introduced as part of the perturbations changed the meaning of the original context. Therefore, to precisely reveal the capability of existing gold standard datasets to benchmark paraphrase understanding, we argue that it is crucial to examine the robustness of MRC systems to paraphrasing the whole context as well.

In this paper, our aim is to evaluate how well current reading comprehension systems generalise to a modified benchmark in which all contexts were paraphrased while preserving the same meaning and thus keeping the same gold standard answer. Different from prior robustness assessment research (Gan and Ng, 2019; Wu et al., 2021a; Yan et al., 2022), we design a pipeline to generate and identify perturbations of MRC examples that demonstrate the lack of robustness of a strong MRC system RoBERTa-large (Liu et al., 2019) to context paraphrasing (see Figure 1 for an example). In doing so, we also underscore the limitation of a large language model to handle the paraphrased contexts. This proposed evaluation framework leads to the construction of a paraphrased test set drawn from the original Stanford Question Answering Dataset v1.1 (SQuAD 1.1) benchmark (Rajpurkar et al., 2016). Results of our experiments show that the performance of five other MRC models except the RoBERTa-large on our created dataset is substantially lower, indicating the transferability of such adversarial attack against MRC models and the insufficiency of the SQuAD 1.1 to benchmark context paraphrasing understanding. Utilising a straightforward training data augmentation approach, we also show the possibility to enhance the robustness of these models in dealing with context paraphrasing. These suggest that there is a need to create gold standard datasets in which context paraphrasing challenges are sufficiently represented.

2 Experiment Setup

MRC Dataset. In this paper, we investigated an extractive English MRC dataset SQuAD 1.1 (Rajpurkar et al., 2016) (License: CC-BY 4.0) due to its simplicity and the fact that it is the dataset on which current MRC models have already achieved superhuman performance, hence allowing us to focus on analysing the robustness of models to context paraphrasing. The statistics for the dataset are reported in Appendix A.

Models. We chose the following models for the task of machine translation and reading comprehension, respectively.

Machine translation: We used the neural translation models provided by OPUS-MT (Tiedemann and Thottingal, 2020) which are based on the popular Marian-Neural Machine Translation framework (Junczys-Dowmunt et al., 2018) pre-trained on the OPUS (Tiedemann, 2012) multilingual corpus.

Reading comprehension: We selected the RoBERTa-large model (Liu et al., 2019) to generate the paraphrased test set mainly due to its impressive performance (93.1% F1) on the original development set of SQuAD 1.1 (Rajpurkar et al., 2016). The process of generating the challenge set also entails the utilisation of a Generative Pre-trained Transformer (GPT) (Brown et al., 2020) series model, specifically GPT-3.5-turbo, through the OpenAI ChatGPT API. In the final evaluation stage, we used multiple strong MRC models including BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2020), SpanBERT (Joshi et al., 2020) and DeBERTa (He et al., 2021), to comprehensively demonstrate the challenge posed by our created dataset. We fine-tuned these pre-trained language models on the training set of SQuAD 1.1 and evaluated them on each of the original and perturbed test sets by making use of HuggingFace's *Transformers* library (Wolf et al., 2020). Model details and the hyperparameters used in model fine-tuning are shown in Appendix B.

3 Context Paraphrasing-Oriented Challenge Set Generation

In this section, we describe our methodology for generating a semantics-preserving context-paraphrased dataset. Four steps are involved in the perturbation pipeline, which are detailed below.

Language	Performance (EM/F1)	
	Original	Paraphrased
Sino-Tibetan		
Chinese	91.07/94.38	80.60/86.05$_{-8.83}$
Indo-European		
Hindi	92.50/95.29	70.89/76.11$_{-20.13}$
Spanish	89.80/93.99	87.75/92.51$_{-1.57}$
French	90.26/94.24	87.20/92.02$_{-2.36}$
Russian	90.69/94.38	85.29/90.27$_{-4.35}$
German	89.95/94.22	87.51/92.29$_{-2.05}$
Italian	90.07/94.13	86.85/91.99$_{-2.27}$
Dutch	89.43/93.98	86.98/92.08$_{-2.02}$
Swedish	89.70/94.04	87.26/92.16$_{-2.0}$
Austronesian		
Indonesian	91.07/94.53	84.39/89.54$_{-5.28}$
Austro-Asiatic		
Vietnamese	91.73/95.06	77.14/82.89$_{-12.8}$
Uralic		
Finnish	91.20/94.62	85.15/90.22$_{-4.65}$

Table 1: The performance (%) of the RoBERTa-large (Liu et al., 2019) on the original and paraphrased test sets generated using 12 pivot languages across five language families. Values in smaller font are changes in F1 (%) relative to the original performance of the model.

3.1 Automatic Context Paraphrasing

We explored paraphrasing the reading passages in the development set of an MRC dataset using a back-translation approach, by which each sentence in the context is translated from a source language (English) to a pivot language and then back to the source language. We identified twelve languages across five language families as the pivot language, informed by their number of speakers (Eberhard et al., 2022) and the performance of their associated pre-trained neural translation models (Tiedemann and Thottingal, 2020). After obtaining the paraphrases, we kept only those with at least one question where all annotated answers can still be found in the paraphrased context. The original contexts of those paraphrases were then extracted from the development set, to keep it aligned with the modified test set and the performance comparable.

3.2 Preliminary Evaluation

As presented in Section 3.1, we generated perturbed test subsets (one for each of the 12 pivot languages) in which contexts were paraphrased using back-translation, and their corresponding original versions. Then, we examined the performance of a strong MRC model, RoBERTa-large (Liu et al., 2019), on these datasets, as demonstrated in Table 1. It can be seen from Table 1 that paraphrasing the contexts using different pivot languages caused various degrees of degradation in terms of the performance of the RoBERTa-large model. Nonetheless, we cannot simply conclude that this indicates the vulnerability of MRC models to the context paraphrasing attack as it is unclear whether these context paragraphs were indeed *paraphrased*, i.e., remain semantically equivalent while lexical/syntactic features were changed. Therefore, we manually verified the validity of the perturbed MRC instances in the next step.

3.3 Human Evaluation

With the aim of studying the lack of robustness of MRC models to context paraphrasing, from each generated perturbed test set, we identified MRC examples on which the RoBERTa-large model (Liu et al., 2019) predicts a wrong answer span whereas it provides the correct answer given the original passage. Afterwards, we randomly sampled 10% examples from each filtered perturbed test set; this resulted in a total of 247 candidate examples, based on which human performance was assessed. A candidate perturbed MRC example has the ability to demonstrate the vulnerability of a model to context paraphrasing, if the model makes a wrong prediction on the paraphrased context paragraph, but a human can answer the question correctly. We refer to such candidates as *suitable* examples. Out of 247 examples, we identified 53 as suitable. The identification process is detailed in Appendix C. In addition, Figure 2 measures the languages contribution of suitable examples within the annotated dataset, from which it is evident that employing Finnish for back-translation/paraphrasing yields the most suitable examples.

3.4 Paraphrased Test Set Generation

While human evaluation enables us to identify suitable MRC instances precisely, it requires significant human annotation effort. Hence, we explored the viability of two different approaches to auto-

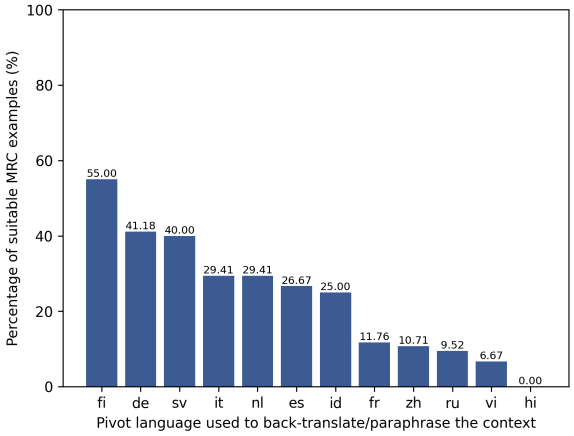

Figure 2: Percentage of suitable MRC examples within the candidate instances generated by using each pivot language, respectively.

matically determine whether a perturbed MRC example is suitable: one based on Machine Learning (ML) techniques and the other employing the GPT-3.5-turbo model. The process and outcomes derived from experimenting with these two methods are detailed in Appendix D and Appendix E. The best-performing model, GPT-3.5-turbo under zero-shot scenario (0.69 precision in predicting suitable example), was then applied on the filtered perturbed instances generated using Finnish, Spanish, Vietnamese, Italian and Swedish, 182 of which were classified as suitable (from 150 original contexts). For multiple paraphrased contexts that correspond to the same original passage, we only kept the perturbed one with the most questions preserved, or in case of a tie, the one with the lowest average question–context lexical overlap (Shinoda et al., 2021). Our final paraphrased test set contains 150 contexts and 158 questions in total. For the purposes of comparison, we also created an *Original* version of the test set keeping only the original passages and questions corresponding to those that were included in the *Paraphrased* version.

4 Results and Discussion

4.1 Evaluation

We assessed the performance of six state-of-the-art MRC models on the newly created challenge set, as shown in Table 2. The table shows that all the evaluated neural language models demonstrated poor generalisation to our generated test set. RoBERTa-large suffered the largest performance drop of 85.07%—this is within our expectation since its errors were used to identify suitable examples. For the other five model architectures, the relative changes were smaller than that of RoBERTa-large, but still very noticeable with over 41% performance decrease. This demonstrates the poor capability of these reading comprehension systems to properly deal with the paraphrased contexts. Apart from RoBERTa and ALBERT, the performance of other four MRC models remained consistent across both original and paraphrased test set, with DeBERTa achieving the highest EM and F1 score, followed by SpanBERT, BERT and DistilBERT. While the performance of ALBERT on the original dataset was slightly lower than that of DeBERTa, it notably outperformed the latter on the paraphrased test set, attaining the highest performance score (54.16 F1). We also found that the consistency in original model performance rankings might not apply to their robustness to context paraphrasing, with the BERT-large model demonstrating the greatest F1 decrease (50.2%) and the ALBERT-xxlarge-v1 exhibiting the smallest performance decline (41.54%).

Model	**Original** (EM/F1)	**Paraphrased** (EM/F1)
RoBERTa-large	100/100	$0/14.93_{-85.07}$
DistilBERT-base	66.46/73.5	$27.22/39.05_{-46.87}$
BERT-large	75.32/81.7	$32.91/40.69_{-50.2}$
SpanBERT-large	77.85/84.72	$32.91/42.62_{-49.69}$
ALBERT-xxlarge-v1	88.61/92.64	$44.3/54.16_{-41.54}$
DeBERTa-large	89.24/93.58	$41.14/49.74_{-46.85}$

Table 2: The performance (%) of the fine-tuned MRC models on the original and the paraphrased test set.

4.2 Error Analysis

To explore the source of model inaccuracies in paraphrased contexts, we manually checked 50 perturbed examples on which the examined MRC models failed and identified three potential sources of model errors. We observed that the paraphrasing of keywords in the sentence that is required to answer the question, along with some other lexical changes, might lead models to provide an incorrect answer (see Figure 9 in Appendix F). Moreover, another source of errors might be the change in the answer sentence structure (see Figure 10 in Appendix F as an example). Paraphrasing other contextual sentences may also inadvertently lead to the generation of incorrect responses by MRC models, particularly when such paraphrases result in keyword overlap with the question. However,

unraveling the sources of these errors in the midst of full-context paraphrasing perturbation remains a complex problem that requires further investigation. Overall, our findings suggest that these high-performing systems might mostly rely on certain words matching between the question and the context to generate the answer, rather than truly understanding the passage. However, we also observed in a small proportion of examples that a mismatch between the answer provided by a model and the gold standard answer, does not necessarily mean that the model's answer is erroneous: in some cases, the semantic meaning of the paraphrased context has changed or the model's answer is arguably correct. This indicates that this work might be underestimating the robustness of the investigated models.

4.3 Robustness Improvement

An intuitive strategy to enhance the models' robustness to context paraphrasing involves exposing them to suitable examples. To this end, we selected 2694 MRC contexts (comprising 12723 questions) from the original SQuAD 1.1 training set (Rajpurkar et al., 2016) and paraphrased them using Finnish, a language that has been demonstrated to be the most effective for generating suitable MRC examples. We then curated the perturbed examples where the answer span still contained within the corresponding paraphrased context, yielding 2459 paraphrased contexts across a total of 8075 questions. The investigated models were then re-trained on the SQuAD 1.1 training set, augmented with these perturbed instances. Table 3 shows their performance on both the original and the paraphrased test sets, before and after re-training.

From Table 3, we can see that on the original test set, apart from the DistilBERT-base, which experienced a slight performance decline in terms of the EM metric, all retrained models demonstrated higher performance than the one trained on the original training set of SQuAD 1.1, though the augmented contexts-paraphrased set contains noises, i.e., those are not suitable examples. These findings showcase the potential for enhancing performance on the original dataset by training models with context-paraphrased MRC examples. On the paraphrased challenge set, for all models expect the DistilBERT-base, re-training with the additional perturbed examples improved the performance and thus their robustness to context paraphrasing. However, for the DistilBERT-base model, exposing it to the paraphrased examples even resulted in a moderate performance drop (7.35% F1), further compromising its robustness in handling context paraphrasing. This might due to the unsuitable examples included in the augmented training set, but also demonstrate the challenging nature of the whole context paraphrasing perturbation on the DistilBERT-base (Sanh et al., 2019).

Model	**Performance (EM/F1)**	
	Before	After
DistilBERT (base)	66.46/73.5	65.19/73.58
	27.22/39.05	24.68/36.18
BERT (large)	75.32/81.7	81.65/85.77
	32.91/40.69	34.18/43.58
SpanBERT (large)	77.85/84.72	83.54/88.96
	32.91/42.62	38.61/48.29
ALBERT (xxlarge-v1)	88.61/92.64	88.61/93.52
	44.3/54.16	46.84/55.78
DeBERTa (large)	89.24/93.58	90.51/94.81
	41.14/49.74	43.04/51.77

Table 3: The performance (%) of MRC systems on the original and the paraphrased test set, before and after re-training. Performance figures displayed in white cells correspond to results obtained on the original test set, whereas the results shown in the shaded areas represent the performance on the generated challenge set.

5 Conclusion

In this paper, we reveal the weaknesses of contemporary reading comprehension systems to context paraphrasing. With the proposed perturbation framework, we generated a paraphrased challenge set, to which six high-performing MRC models generalise poorly. We also demonstrate that a training data augmentation approach can enhance the robustness of the majority of models when exposed to the paraphrased contexts. This informs us that to equip models with context paraphrasing understanding ability, there is a need to create benchmarks in which this reasoning challenge is precisely represented. Future work will include the design of better techniques to remove the noise existing in the challenge set and the optimise of the perturbation pipeline so that it can be generalisable to more challenging datasets.

Limitations

In this work, our annotated gold dataset might contain potentially debatable instances of suitable MRC examples. To address this concern, there is a pressing need for the establishment of theoretical foundations which clearly define *human answerable* under the context-paraphrasing oriented perturbations and other types of perturbations. Building upon this, research efforts are needed to evaluate and enhance the precision of automatic approaches for identifying suitable examples, enabling precise assessment of models robustness against test-time perturbations. Further, there is potential to design better document-level paraphrasing methods and expand this study to include other sophisticated MRC datasets and diverse NLU tasks.

Ethics Statement

The SQuAD 1.1 benchmark and models utilized in this paper are all publicly available. During the phase of human annotation, explicit instructions regarding the annotation task were provided to all annotators, and they were informed about the intended use of their annotations. Only their responses to the questions were collected, and no sensitive information was gathered. Consequently, we do not anticipate any ethical concerns arising from our work.

Acknowledgements

The authors express the gratitude to the anonymous reviewers from the ACL Rolling Review April and June 2023 for their invaluable suggestions. We would also like to thank Michael White and Jing Wang in annotating the perturbed MRC examples. Part of the experiments were conducted with the support of the Computational Shared Facility at The University of Manchester, for which the authors are grateful. This work was supported by the University of Manchester Department of Computer Science Kilburn Scholarship.

References

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.

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

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

Scott A. Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1):14–27.

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.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2022. Ethnologue: Languages of the World. Available at `https://www.ethnologue.com`.

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.

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

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. 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*.

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.

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, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. 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.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2021. Can question generation debias question answering models? a case study on question–context lexical overlap. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 63–72, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2023. Which shortcut solution do question answering models prefer to learn? *Proceedings of the AAAI Conference on Artificial Intelligence*.

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.

Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. 2022. What makes reading comprehension questions difficult? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6951–6971, Dublin, Ireland. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. 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, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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. 2021a. 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. 2021b. Is the understanding of explicit discourse relations required in machine reading comprehension? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3565–3579, Online. 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.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models.

A Dataset Statistics

Table 4 presents the number of contexts and questions contained within the SQuAD 1.1 training and development set, respectively.

	Training	Development
Context	18,896	2,067
Question	87,599	10,570

Table 4: Number of contexts and questions in the SQuAD 1.1 training and development sets (Rajpurkar et al., 2016).

B Hyperparameters of the Neural Reading Comprehension Models

Table 5 shows the hyperparameters used to fine-tune the pre-trained MRC models in this work. We utilised 2 16GB Nvidia v100 GPUs to fine-tune and evaluate each model.

$\text{Model}_{parameters(M)}$	d	b	lr	ep
RoBERTa-large$_{(355)}$	384	8	3e-5	2.0
DistilBERT-base$_{(65)}$	384	8	3e-5	3.0
BERT-large$_{(340)}$	384	8	3e-5	2.0
SpanBERT-large$_{(340)}$	512	4	2e-5	4.0
ALBERT-xxlarge-v1$_{(223)}$	384	4	3e-5	2.0
DeBERTa-large$_{(350)}$	384	4	3e-6	3.0

Table 5: The hyperparameters used to fine-tune each pre-trained MRC model (with its number of parameters). 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.

C Human Annotation

This Appendix details the process of manually identifying the perturbed MRC examples that are suitable for context paraphrasing oriented robustness assessment. A total of three human annotators were involved in this task, including the first author of this paper. Prior to starting the annotation task, we asked all annotators to check a few examples and report the average time they spent for annotating each example. Based on this, we paid the annotators for their work by offering them coupons with a value of 20 pence for each example they annotate.

For the randomly sampled 247 candidate instances, we first asked two annotators to answer each question based on the corresponding paraphrased context, respectively. The annotators were required to select the shortest continuous span in the paraphrased context that answered the question only if they are confident that the paraphrased context still makes it possible to answer the associated question and were allowed to leave the answer as blank if the question is not answerable anymore. Full text of instruction given to the annotators can be seen in Figure 3. Afterwards, for each example, we measured the correctness of the answer span provided by each annotator through comparing it with the ground truth answers, respectively , and labelled the example as suitable or unsuitable based on the criteria described in Section 3.3. To conduct a precise analysis, we manually checked all examples with the answer span given by the annotator(s) does not exact match any of the ground truth answers and decided the correctness of the answer by taking into account the corresponding context and the question as well. Figure 4 demonstrates one such example. We then measured the inter-annotator agreement by computing the Cohen's kappa coefficient (Cohen, 1960), which is around 0.48. This might indicate that there exists moderate discrepancies between the two annotators concerning the answerability of the questions predicated on the contexts that have been paraphrased. Finally, we presented the examples on which the two annotators share a disagreement to the third annotator and provided them the label that agreed

by the majority of annotators. This yielded a total of 66 suitable examples. From the identified 66 examples, we further manually eliminated 13 wherein the prediction of the RoBERTa-large (Liu et al., 2019) could be reasonably deemed accurate, thus rendering them unsuitable for the robustness assessment (see Figure 5 as an example). Our final annotated dataset contains 53 (out of 247) suitable examples. In an effort to curtail potential bias, all annotators were solely provided with the paraphrased context and the corresponding question for their examination.

Thanks for contributing to this project! Your task is to read each given context and answer a question about it. We will compare the answer you provide with the ground truth answers to determine the human answerability of the question, and then screen out the examples that are suitable for the robustness assessment of reading comprehension systems. When you are answering the questions:
(1) If you meet a question that you truly think you can answer it based on the given context, then select the shortest continuous span in the context as the answer.
(2) If you meet a question that is completely unanswerable, leave the answer as blank.

Figure 3: Instructions for the annotation task.

Context: [. . .] The production of major food staples such as corn is subject to sharp weather-related fluctuations. [. . .]

Paraphrased Context: [. . .] The production of staple foods, such as maize, is affected by severe weather-related fluctuations. [. . .]

Question: What can cause fluctuations in the production of corn?

Ground Truth Answers:
weather-related fluctuations, weather-related, weather
Prediction Under Context Paraphrasing:
Human Annotator 1:
severe weather
Human Annotator 2:
severe weather-related

Figure 4: An instance requiring human effort for the validation of answer accuracy. Both answer spans provided by the two annotators are considered correct, despite yielding an EM score of 0.

Context: Kenya [. . .], officially the Republic of Kenya, is a country in Africa and a founding member of the East African Community (EAC). Its capital and largest city is Nairobi. Kenya's territory lies on the equator and overlies the East African Rift covering a diverse and expansive terrain that extends roughly from Lake Victoria to Lake Turkana (formerly called Lake Rudolf) and further south-east to the Indian Ocean. [. . .]

Paraphrased Context: Kenya (Kenya: "Kenya") is a country in Africa and one of the founding members of the East African Community (EAC). The capital and largest city is Nairobi. The area of Kenya lies on the equator and survives the East African Rift which covers a diverse and vast area that stretches roughly from Lake Victoria to Lake Turkana (formerly Lake Rudolf) and further south-eastern to the Indian Ocean. [. . .]

Question: Where is Kenya located?

Ground Truth Answers:
Africa, in Africa
RoBERTa-large's Prediction Under Context Paraphrasing: on the equator

Figure 5: A perturbed example that is not suitable for the robustness assessment since the answer span offered by the RoBERTa-large model (Liu et al., 2019) is reasonably accurate, albeit not an exact match for any of the ground truth answers.

D Automated Identification of Suitable MRC Instances

To circumvent the substantial effort required for manual annotation, we attempted to automatically classify whether a perturbed reading comprehension example is qualified to demonstrate the lack of robustness of MRC models to context paraphrasing. In the following, we elaborate on the two approaches undertaken and present the empirical results derived from these experiments.

ML-based Approach: We trained and evaluated multiple classifiers on our 247 annotated examples with 129 input features that were calculated by TAACO (Crossley et al., 2019), a tool that measures various linguistic features of the passage such as lexical density and adjacent sentence overlap. The designed classification pipeline involves data standardisation, features selection and random oversampling. Hyperparameter tuning was carried out to

determine the optimal configuration. The obtained best-performing model, Random Forest (with 40 selected features), only achieved 0.39 precision in predicting suitable example, which implies that those extracted features might not sufficient to represent this challenging task. Therefore, we shifted our attention to the GPT series models, given their exceptional efficacy in transforming many tasks into generative tasks.

GPT Series Models: Compared to traditional ML methods, GPT series models offer the advantage of not requiring the construction of linguistic features, thereby simplifying the approach to automatically classify suitable MRC examples. Drawing upon the human annotation process described in Appendix C, we first manually constructed the zero-shot prompt encompassing the paraphrased context, question, ground truth answers, the answer span given by the RoBERTa-large (Liu et al., 2019), and tasked the model to generate binary output (0 or 1) to indicate whether an example is suitable for robustness assessment, adhering to a predefined set of decision rules. We also experimented with the few-shot prompt by adding three randomly selected in-context exemplars of input-label pairs (demonstrations) (Brown et al., 2020) in the zero-shot prompt. Under both zero-shot and few-shot scenarios, we further investigated the use of the Chain-of-Thought (CoT) (Wei et al., 2022) by adding "let's think step by step" and CoT demonstrations in the corresponding prompt, respectively. The templates for the four prompting strategies are shown in Appendix G. In order to mitigate the influence of prior dialogues, each request was sent individually to produce the corresponding response. When processing the responses, especially under the zero-shot CoT and few-shot CoT scenarios, we only consider an example as suitable if its response includes a solid explanation and judgement. Labels generated by the model under the four distinct test configurations were subsequently compared with the gold labels annotated by human evaluators, respectively. The results are shown in Table 6.

It can be seen from Table 6 that on the precision of predicting suitable MRC example, prompting under the zero-shot scenario provides the best result, which is 0.41. Surprisingly, the incorporation of demonstrations and the adoption of the CoT prompting considerably attenuate model performance, a finding that deviates from existing literature asserting enhancements in performance

Prompting Method	Precision
Zero-shot	0.41
Zero-shot CoT	0.23
Few-shot	0.26
Few-shot CoT	0.28

Table 6: Precision of the GPT-3.5-turbo model in predicting suitable example using four different prompting methods.

across many NLU tasks with the inclusion of in-context demonstrations (Brown et al., 2020) and the CoT method (Wei et al., 2022). The observed unsatisfactory performance could potentially be attributed to two factors: (1) The ambiguity inherent to the task of automated identification of suitable MRC example, as viewed from the dataset annotation perspective. As indicated in Appendix C, a moderate level of disagreement was even observed between two human annotators in determining whether a question is indeed answerable based on the paraphrased context, with an inter-annotator agreement score of 0.48. This suggests that our annotated set of 247 examples might contain contentious cases, thereby rendering the task notably challenging for the model. (2) From the model's perspective, we investigated potential reasons for performance degradation following the adoption of in-context examples and CoT by manually scrutinizing some responses under these testing conditions. Our findings reveal that despite guidance from demonstrations and CoT, the model frequently produces reasoning that contradicts the predicted label or even generates hallucinations. For instance, under the zero-shot CoT scenario, model produces response like "*This example is suitable for robustness assessment. The ground truth answers (GTAs) and RoBERTa's answer A are different, indicating that there is potential for the model to make mistakes. Therefore, it is important to test the model's robustness by presenting it with similar but slightly different contexts and questions to ensure that it can generalize well and provide accurate answers.*", which even not relevant to the task. While we acknowledge that there exists scope to improve the prompts used in this work, it remains evident that the GPT-3.5-turbo model, despite its significant accomplishments in some NLU tasks, still falls short of attaining human-level language comprehension.

Though the performance of the obtained best model, i.e., GPT-3.5-turbo under the zero-shot sce-

nario, is not satisfactory, we attempted to measure its precision in predicting suitable example from the candidate perturbed instances generated by using each pivot language, respectively, as shown in Table 7. From Table 7, we can see that in classifying perturbed examples paraphrased using Finnish, the model exhibits flawless performance, achieving a precision score of 1.0. In contrast, for other languages, such as Russian, Chinese and Indonesian, the precision score is notably low or even zero. Therefore, to generate our paraphrased challenge set, we restricted model's application to filtered perturbed examples produced using Finnish, Spanish, Vietnamese, Italian and Swedish (with precision greater than 0.5), which yielded a final precision score of 0.69, while excluding instances generated using other languages.

Pivot Language(s)	Precision
Finnish	1.0
Spanish	0.8
Vietnamese	0.67
Italian, Swedish	0.5
German	0.33
Dutch, Russian	0.25
Chinese	0.16
Hindi, Indonesian, French	0

Table 7: Precision of the GPT-3.5-turbo in predicting suitable example from the candidate perturbed instances generated using each of the 12 pivot languages.

E Analysis of Disparities in Adversarial Examples Perception: GPT-3.5-turbo vs. Human

The zero-shot prompt, as designed in Appendix D, directly solicits a binary response from the GPT-3.5-turbo model concerning the suitability of a perturbed MRC example for robustness assessment. To validate the stability of the obtained performance (0.41 precision) and also examine the divergence in the perception of whole context-paraphrased adversarial examples between the GPT-3.5-turbo model and human observers, we conducted further experiments by directly asking the model to extract the shortest continuous span as the answer given the paraphrased context and the question. We utilised the prompting method based on both instruction and opinion (Zhou et al., 2023) to improve the faithfulness of the model to paraphrased context when formulating responses, thereby precluding the use of its parametric knowledge to a great extent. Additionally, an "I do not know" option was allowed to encourage the model to abstain from providing the answer if the paraphrased context does not make it possible to answer the question anymore. Figure 6 demonstrates the used prompt template.

Instruction: read the given information and answer the corresponding question. The output should only be the shortest continuous span from the context and should not include any explanation. Output "I do not know" if the context makes it impossible to answer the corresponding question.
Bob said, "context"
Q: question in Bob's opinion based on the given text?

Figure 6: An instruction-opinion based prompt template (Zhou et al., 2023).

Afterwards, we determined the accuracy of each response generated by the GPT-3.5-turbo model and assigned a binary label (1 or 0) to signify its appropriateness for robustness assessment. We then compared the obtained results with the gold standard labels of the annotated dataset version containing 66 suitable examples (see Appendix C), as our provided prompt does not require the model to consider the correctness of the prediction made by the RoBERTa-large (Liu et al., 2019). Experimental results revealed that the GPT-3.5-turbo model maintained a consistent 0.41 precision score in predicting suitable MRC example, thereby suggesting its stability on this task to some extent. Figure 7 and Figure 8 demonstrate a failure case of the GPT-3.5-turbo model on the suitable example classification, respectively. In Figure 7, the model is still able to extract the correct answer span "John Sutcliffe" from the paraphrased context, though both human annotators deem that the question is not answerable. On the contrary, as can be seen from Figure 8, while human annotators can get the answer correct, the model abstains from answering the question and thus generates "I do not know" as the answer. These findings suggest the existence of a significant gap between the GPT-3.5-turbo model and human performance in discerning whole context paraphrasing oriented textual attacks and the GPT-3.5-turbo is still substantially distant

from achieving human-level NLU capability.

Context: [. . .] The game was called by ESPN Deportes' Monday Night Football commentary crew of Alvaro Martin and Raul Allegre, and sideline reporter John Sutcliffe. [. . .]
Paraphrased Context: [. . .] The game was called by ESPN Deportes' Monday Night Football comment crew Alvaro Martin and Raul Allegre, and side EPOR John Sutcliffe. [. . .]
Question: Who was the ESPN Deportes sideline commentator for Super Bowl 50?
Prediction Under Context Paraphrasing: **Human Annotators:** Unanswerable **GPT-3.5-turbo:** John Sutcliffe

Figure 7: Demonstration of a failure case of the GPT-3.5-turbo model in predicting suitable example. While both human annotators deem that the question is not answerable over the paraphrased context, the model still provides the correct answer span.

Context: Sudbury model democratic schools claim that popularly based authority can maintain order more effectively than dictatorial authority for governments and schools alike. They also claim that in these schools the preservation of public order is easier and more efficient than anywhere else. [. . .]
Paraphrased Context: Model schools in Sudbury argue that popular authority can maintain order more effectively than dictatorial authority for governments and schools. They also claim that, in these schools, the preservation of public order is easier and more effective than anywhere else. [. . .]
Question: In addition to schools, where else is popularly based authority effective?
Prediction Under Context Paraphrasing: **Human Annotators:** governments **GPT-3.5-turbo:** I do not know

Figure 8: Illustration of the robustness deficiency of the GPT-3.5-turbo model to whole context paraphrasing. The model was unable to generate the correct answer span, despite both human annotators supplying the accurate response.

F Suitable Examples Demonstration

We present two perturbed examples from the constructed challenge set on which the MRC models demonstrated unsatisfactory generalisation, as shown in Figure 9 and Figure 10, respectively.

Paragraph: The game's media day, which was typically held on the Tuesday afternoon prior to the game, was moved to the Monday evening and re-branded as Super Bowl Opening Night. [. . .]
Paraphrased Paragraph: The video day of the game, which was usually held on Tuesday afternoon before playing games, was moved to Monday night and was reset as Super Bowl Opening Night. [. . .]
Question: What day of the week was Media Day held on for Super Bowl 50?
Original Prediction: Monday **Prediction Under Adversary:** Tuesday

Figure 9: A perturbed example primarily involves lexical changes, ultimately leading the model to provide the wrong answer.

Paragraph: A method to lessen the magnitude of this heating and cooling was invented in 1804 by British engineer Arthur Woolf, who patented his Woolf high-pressure compound engine in 1805. [. . .]
Paraphrased Paragraph: British engineer Arthur Woolf, who patented Woolf's high pressure engine in 1805, invented in 1804 a method to reduce the volume of this heating and cooling. [. . .]
Question: What nationality was Arthur Woolf?
Original Prediction: British **Prediction Under Adversary:** engineer

Figure 10: Illustration of the brittleness of MRC systems when dealing with a syntactic form changed context.

G Templates for Various Prompting Strategies

Figure 11 illustrates the diverse templates employed to prompt the GPT-3.5-turbo model to classify the suitable perturbed MRC example.

Instructions: Given an example which contains a context, question, ground truth answers (GTAs) and RoBERTa's answer A, decide whether it is suitable for robustness assessment by choosing one of the following options: '0': A is reasonably correct or A is wrong and you cannot correctly answer the question purely relying on the context as well. '1': A is wrong but you can correctly answer the question purely relying on the context.

zero-shot:
[Instructions] Generate either '0' or '1', do not include the explanation.
Context: [context] Question: [question] GTAs: [GTAs] A: A

zero-shot CoT:
[Instructions]
Context: [context] Question: [question] GTAs: [GTAs] A: A
Let's think step by step and then generate the response ([0] or [1]):

few-shot:
[Instructions] Generate either '0' or '1', do not include the explanation.

Example:
Context: Model schools in Sudbury argue that popular authority can maintain order more effectively than dictatorial authority for governments and schools. [...]
Question: In addition to schools, where else is popularly based authority effective?
GTAs: ['governments']
A: governments and schools. They also claim that, in these schools, the preservation of public order is easier and more effective than anywhere else.
Response: 1

Example:
Context: [context]
Question: [question]
GTAs: [GTAs]
A: A

few-shot CoT:
[Instructions]

Example:
Context: Model schools in Sudbury argue that popular authority can maintain order more effectively than dictatorial authority for governments and schools. [...]
Question: In addition to schools, where else is popularly based authority effective?
GTAs: ['governments']
A: governments and schools. They also claim that, in these schools, the preservation of public order is easier and more effective than anywhere else.
Response: Firstly, compare RoBERTa's answer A with GTAs. Since *governments and schools. They also claim that, in these schools, the preservation of public order is easier and more effective than anywhere else.* is wrong, then there is a need to thoroughly check the context and question. Since the context provides sufficient information to enable us to get the answer correct, the response is 1.

Example:
Context: [context]
Question: [question]
GTAs: [GTAs]
A: A

Figure 11: Prompt templates provided to the GPT-3.5-turbo model. Due to space limitations, we only show one in-context input-label pair in the few-shot and few-shot CoT template.