

# Towards Confident Machine Reading Comprehension

Rishav Chakravarti\*

AWS AI Labs  
New York City, NY  
chakrris@amazon.com

Avirup Sil†

IBM Research AI  
Yorktown Heights, NY  
avi@us.ibm.com

## Abstract

There has been considerable progress on academic benchmarks for the Reading Comprehension (RC) task with State-of-the-Art models closing the gap with human performance on extractive question answering. Datasets such as SQuAD 2.0 & NQ have also introduced an auxiliary task requiring models to predict when a question has no answer in the text. However, in production settings, it is also necessary to provide confidence estimates for the *performance* of the underlying RC model at both answer extraction and “answerability” detection. We propose a novel post-prediction confidence estimation model, which we call MR.C (short for Mr. Confident), that can be trained to improve a system’s ability to refrain from making incorrect predictions with improvements of up to 4 points as measured by Area Under the Curve (AUC) scores. MR.C can benefit from a novel white-box feature that leverages the underlying RC model’s gradients. Performance prediction is particularly important in cases of domain shift (as measured by training RC models on SQuAD 2.0 and evaluating on NQ), where MR.C not only improves AUC, but also traditional answerability prediction (as measured by a 5 point improvement in F1).

## 1 Introduction

The reading comprehension (RC) task require models to extract or generate answers to questions about an input piece of text. Particularly since the advent of benchmark datasets such as SQuAD (Rajpurkar et al., 2016), transfer learning models that leverage large pre-trained language models (LM) like BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2019) have demonstrated high performance, even rivaling “human” performance.

\*Work completed while at IBM Research AI

†Corresponding author

### Contradictory Context

**Context:** Concrete hardens as a result of the chemical reaction between cement and water (known as hydration...For every pound (or kilogram or any unit of weight) of cement, **about 0.35 pounds** (or 0.35 kg or corresponding unit) of water is needed... However, **a mix with a ratio of 0.35 may not mix thoroughly**, and...ratios of **0.45 to 0.60 are more typically used**.

**Question:** minimum required water cement ratio for a workable concrete is

**Base RC Prediction:** 0.35

**Base RC Score:** 0.81

**MR.C Score:** 0.19

**Query Embedding Gradient Highlighting:**

minimum required water cement ratio  
for a workable concrete is

### RC Model Ignores Part of Question

**Context:** The **Inca Empire** at its greatest extent Capital Cusco (1438–1533)...Religion Inca religion Government Divine, **absolute monarchy**...

**Question:** what government structure did the aztec and inca have in common

**Base RC Prediction:** absolute monarchy

**Base RC Score:** 0.94

**MR.C Score:** 0.41

**Query Embedding Gradient Highlighting:**

what government structure did the  
aztec and inca have in common

Figure 1: Examples from the NQ dataset of bad answers from Base RC Model (in our case ALBERT<sub>QA</sub>) nonetheless produces high confidence scores. Our MR.C model produces lower confidence scores in these cases facilitating better thresholding.

Despite this progress, the reliability of NLP models continues to be a concern with recent work demonstrating a lack of robustness to adversarial perturbations (Jia and Liang, 2017; Wang and Bansal, 2018; Wallace et al., 2019) as well as slight domain drift between training and inference (Niu and Bansal, 2019). Despite these challenges, RC technologies are being adopted in industry settings (e.g. Amazon Kendra<sup>1</sup> offers users the ability to ex-

<sup>1</sup><https://aws.amazon.com/kendra>

tract answers from their own corpora as part of its search offerings) where bad predictions can have adverse implications both in terms of reputation as well as actual harm (Hern, 2017; Klar, 2019).

Recent benchmarks such as SQuAD 2.0 (SQ2) (Rajpurkar et al., 2018) and Natural Questions (NQ) (Kwiatkowski et al., 2019) have attempted to address reliability by introducing “unanswerable” questions into the datasets so as to force RC models to generate “no answer” predictions. SQ2 relies on crowd workers to generate both types of questions based on Wikipedia paragraphs while NQ organically collects questions from search logs and uses crowdworkers to annotate them as answerable or not based on Wikipedia articles. However, this modification to the RC task only addresses the model’s ability to recognize bad inputs (the traditional “out-of-domain” question detection (Jia and Xie, 2020; Tan et al., 2019)). While this is important, in production settings, it is also necessary to understand when the model performs poorly even in cases when an answer exists. See Fig. 1 highlighting examples from the NQ dataset where a (near) State-of-the-Art (SOTA) RC model fails to extract the correct answer, but makes a prediction nonetheless (with a high confidence score). As we discuss in section 4, such scenarios are particularly frequent when slight domain shifts take place (a common occurrence in real-life production scenarios).

This paper carries out a novel preliminary study on performance prediction where near SOTA RC models are evaluated on their ability to refrain from making incorrect predictions (both in the case of unanswerable questions as well as in the case where the model simply fails to make the right prediction despite its presence in the text). Using the SQ2 and NQ datasets, we find that:

1. A novel post-prediction confidence estimation model, which we call MR.C (short for Mr. Confident), can be trained to improve a system’s ability to refrain from making incorrect predictions with improvements of up to 4 points as measured by Area Under the Curve (AUC) scores.
2. MR.C can benefit from a novel white-box feature that leverages the underlying RC model’s gradients.
3. Performance prediction is particularly important in cases of domain shift (as measured by training RC models on SQ2 and evaluating on

NQ), where MR.C not only improves AUC, but also traditional answerability prediction (as measured by a 5 point improvement in F1).

## 2 Related Work

**Answer Re-ranking:** Wang et al. (2017) aggregate evidences from different passages for open-domain QA and Kratzwald et al. (2019) uses a combination of retrieval and comprehension features that are directly extracted from the QA pipeline. Nogueira and Cho (2019) perform passage re-ranking using BERT which is another post-prediction task, but can suffer from the same challenge in production settings where the re-rank score is not necessarily informative for the purposes of thresholding “good” answers from “bad” ones.

**Answer Verification:** Tan et al. (2018); Hu et al. (2019); Zhang et al. (2020a,b) perform answer validation and Peñas et al. (2007) adds an extra “verifier” component to decide whether the predicted answer is entailed by the input snippets. Back et al. (2020) perform requirement inspection by computing attention-based satisfaction score to compare question and candidate answer embeddings. Despite their similarity, these approaches continue to focus on discriminating between “answerable” and “unanswerable” questions without directly estimating the underlying RC model’s prediction performance (regardless of answerability with respect to the context). This work’s focus on performance prediction (using metrics such as AUC) seems complimentary in nature to improvements in unanswerability detection.

**Query Performance Prediction (QPP):** Carmel and Yom-Tov (2010); He and Ounis (2006); Tao and Wu (2014) predict retrieval performance of document retrieval which aligns with our objective. However, as shown in our novel use of embedding gradients, there is room to introduce features and model types that are more tailored to the RC setting, which is not studied in the prior work.

## 3 Method

### 3.1 Base RC Model Overview

To fulfill our objective to build a confident RC system, we start with a SOTA RC model as our base. Specifically, given a token sequence ( $\mathbf{X}$ ) consisting of a question and a passage and special markers *e.g.* the  $[CLS]$  token for answerability classification, the base RC model trains classifiers to predict the begin and end of answers

spans as follows:  $\alpha_b = \text{softmax}(\mathbf{W}_1 \mathbf{H})$  and  $\alpha_e = \text{softmax}(\mathbf{W}_2 \mathbf{H})$  where  $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{1 \times D}$ , where  $\mathbf{H}$  is the contextualized representation of  $\mathbf{X}$  provided by a deep Transformer (Vaswani et al., 2017) based language model (LM) and  $D$  is the LM’s output embedding dimension. An extra dense layer ( $\mathbf{W}_3 \in \mathbb{R}^{5 \times D}$ ) is added which operates only on the contextualized representation of the  $[CLS]$  token to produce a likelihood prediction for additional answer types (required for NQ)<sup>2</sup>:  $\alpha_i^{[CLS]} = \text{softmax}(\mathbf{W}_3 \mathbf{h}^{[CLS]})$ . We choose ALBERT (xxlarge v2) (Lan et al., 2019) as our LM and build our RC model, henceforth ALBERT<sub>QA</sub>, following (Zhang et al., 2020b) for SQ2 and (Pan et al., 2019; Alberti et al., 2019) for NQ<sup>3</sup>. For more details, we direct interested readers to their papers and the appendix for hyperparameter settings.

### 3.2 Confidence Estimation Model Overview

**Learning Algorithm:** We introduce MR.C, a post prediction confidence estimation model, that utilizes a Gradient Boosted Machine (GBM) (Friedman, 2001) to learn an ensemble of weak learners (regression trees). The objective function uses logistic regression loss to predict a binary label,  $\mathbf{Y}$ , indicating whether the max scoring answer span from the base ALBERT<sub>QA</sub> model correctly answers the question as per the official evaluation script’s exact match criteria (a *NULLSPAN* indicator is expected for “unanswerable” questions).

To derive training labels, we train ALBERT<sub>QA</sub> on a 90% random split of the training data and generate predictions on the remaining 10%. Both SQ2 and NQ are large datasets ( $\sim 130K$  query-passage pairs and  $\sim 300K$  query-article pairs respectively) so there is not a statistically significant deterioration going from a 100% of the data to 90% of the data. However, k-fold cross validation can be used to generate additional training data for MR.C.

**Input Features:** The following “grey box” features are used as input to MR.C:

1. **TopK-BeginLikelihood:**  $\alpha_b$  for each of the top  $k$  spans.
2. **TopK-EndLikelihood:**  $\alpha_e$  for each of the top  $k$  spans.
3. **Offset-Overlap-With-Top1:** the offset based F1 score between the predicted top 1 answer span and each of the remaining  $k - 1$  spans.

4. **Query-IDF:** compute 4 features  $[\min, \max, \text{mean}, \text{skew}]$  of the inverse document frequency scores for each of the input query tokens  $q$  as computed on an English Wikipedia corpus.
5. **NoAnswer-Likelihood:** Compute  $(\alpha_b^{[CLS]} + \alpha_e^{[CLS]})$ . Typically, RC systems (Devlin et al., 2019) use this score to predict a “no-answer” as opposed to an answer string from the passage.
6. **AnswerType-Likelihood:** Compute  $\alpha_l$  for all possible answer types (computed for NQ only).

In addition, we derive a set of novel “white box” features: **Query-Embedding-Gradients** (QEG). We first use ALBERT<sub>QA</sub>’s top scoring  $\alpha_b^{t_i} + \alpha_e^{t_j}$  offsets as the target labels to compute cross entropy loss. This loss is back propagated to compute the gradients for each of the input query token embeddings (i.e. the input of the underlying ALBERT LM)<sup>4</sup>. Finally, we compute  $[\min, \max, \text{mean}, \text{skew}]$  over each of the token embedding gradient norms.

**Intuition:** Features 1–4 are inspired by previous work in QPP (Carmel and Yom-Tov, 2010): 1–3 derive signals from the top answer and potential other answers as a post retrieval scoring mechanism, whereas 4 derives inspiration from specificity which tend to favor specific queries over general ones. Features 5 & 6 track “unanswerability” prediction. Finally, QEG looks at how well aligned the underlying query terms were with the context as seen by ALBERT<sub>QA</sub>.

## 4 Experiments

**Dataset + Conditions:** We evaluate MR.C on these datasets under three different training conditions:

1. **Train<sub>SQ2</sub>-Test<sub>SQ2</sub>:** Train on SQ2 + evaluate on SQ2.
2. **Train<sub>SQ2</sub>-Test<sub>NQ</sub>:** Train on SQ2 + evaluate on NQ Short Answer<sup>5</sup>. We use this as a measure model robustness since both tasks are fairly similar.
3. **Train<sub>NQ</sub>-Test<sub>NQ</sub>:** Train on NQ + evaluate on NQ: Short and Long Answers

<sup>2</sup>We only model long, short and null following prior work.

<sup>3</sup>At time of writing this was published single model SOTA.

<sup>4</sup>We take the  $L^2$  norm to compute a single value per token

<sup>5</sup>SQ2 does not have long answer selection, so we provide the model with the oracle long answer paragraphs.

Train on SQ2	SQ2		NQ	
	F1	AUC	F1	AUC
(Zhang et al., 2020b)	88.8	-	-	-
ALBERT <sub>QA</sub>	89.1	81.1	62.0	55.4
+ MR.C	88.7	<b>85.0</b>	<b>68.2</b>	<b>65.9</b>

  

Train on NQ	Long Answer		Short Answer	
	F1	AUC	F1	AUC
(Pan et al., 2019)	68.2	-	57.2	-
ALBERT <sub>QA</sub>	71.0	85.4	60.1	92.5
+ MR.C	<b>72.8</b>	<b>88.0</b>	59.8	92.6

Table 1: Bold denotes statistically significant differences. We also compare against the prior published SOTA. F1 scores are calculated using the official evaluation scripts for SQ2 and NQ.

**Evaluation Metric:** Rather than focusing purely on the official F1 measure evaluated at an optimally chosen threshold to balance precision on “answerable” questions while limiting recall of “unanswerable” questions, we evaluate using area under the receiver operating characteristics (ROC) curve (AUC) measuring model prediction correctness (regardless of “answerability”) to summarize performance across a variety of threshold settings (Jin Huang and Ling, 2005). In all cases, we evaluate on the official dev sets since this requires question level model performance metrics which are inaccessible for the test sets<sup>6</sup>.

**Results:** Table 1 show gains in AUC over ALBERT<sub>QA</sub> (denoted as Base in the tables) across all train and dev configurations with the largest improvement being in the **Train<sub>SQ2</sub>-Test<sub>NQ</sub>** setting suggesting that MR.C generalizes better (perhaps due to more robust features or simpler modeling assumptions). The AUC increases are statistically significant ( $\alpha = 0.01$  based on a bootstrap randomization test) with the exception of the Short Answer setting in **Train<sub>NQ</sub>-Test<sub>NQ</sub>**. The NQ Short Answer subset is skewed towards “unanswerable” questions and it seems to do a reasonably good job at learning to identify these questions. However, it is still failing to detect prediction errors on “answerable” questions (we provide a breakdown in the appendix where we see that the AUC on the answerable subset is 2.7 points higher for MR.C). This suggests future work to extract features which better leverage the NQ trained ALBERT<sub>QA</sub> model’s ability to detect unanswerable questions.

In addition, the official NQ script’s exact match is stricter than that of SQ2. SQ2 allows slight de-

viations (e.g. casing) and correct tokens can be predicted from any context within the passage. During evaluation, multiple annotations per example are available in the NQ dev set (so the answer can match any of the variations), but the train set (used to provide labels for MR.C training) only contains a single annotation per example. As future work, we intend to explore softer matches to encourage MR.C’s hill climbing during the training process.

There is a slight drop in the base F1 measurement (which only looks at “un-answerability” and not on performance prediction) on **Train<sub>SQ2</sub>-Test<sub>SQ2</sub>** as well as the Short Answer portion of **Train<sub>NQ</sub>-Test<sub>NQ</sub>**. However, neither of these drops are statistically significant and, in fact, there is a dramatic and statistically significant F1 increase in the **Train<sub>SQ2</sub>-Test<sub>NQ</sub>** setting. So overall, MR.C tends to be a more robust decider for whether to trust the underlying ALBERT<sub>QA</sub> model prediction.

**Feature analysis:** We also analyze the informativeness of the input features for MR.C and find that the *mean QEG* feature is the most influential feature in the **Train<sub>SQ2</sub>-Test<sub>NQ</sub>** and **Train<sub>SQ2</sub>-Test<sub>SQ2</sub>** (as measured by the average decrease in mean squared error over tree nodes involving this feature in the GBM). Refer to the appendix for a full list of features sorted by their relative influence.

During manual error analysis on a random sample, we also find that this feature appears to provide a useful visualization of the errors with respect to the query terms *e.g.* Fig 1 shows the query tokens highlighted based on their *QEG* values. The terms with the largest gradients do seem like the critical ones which are least aligned with the evidence in the context so may provide useful insight to the reader.

## 5 Conclusion

Making errors in academic benchmarks results in a lower F1 score but making errors *confidently* in production systems is, at best, embarrassing (Klar, 2019) and, at worst, harmful (Hern, 2017). We propose MR.C, a post prediction confidence estimation model that maintains the base system accuracy, while providing statistically significantly better judgments for the RC model prediction’s accuracy. Previous work in RC has only addressed either re-ranking or detecting “un-answerable” questions rather than directly modeling the prediction error of the base RC system: we hope this novel

<sup>6</sup>Leaderboard results show that dev generalizes to test



work in RC will spur a future research direction and lead to more confident RC systems.

## References

- Chris Alberti, Kenton Lee, and Michael Collins. 2019. [A BERT baseline for the natural questions](#). *arXiv preprint arXiv:1901.08634*, pages 1–4.
- Seohyun Back, Sai Chetan Chinthakindi, Akhil Kedia, Haejun Lee, and Jaegul Choo. 2020. Neurquri: Neural question requirement inspector for answerability prediction in machine reading comprehension. *ICLR*.
- David Carmel and Elad Yom-Tov. 2010. *Estimating the query difficulty for information retrieval*. Morgan & Claypool.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- Jerome H. Friedman. 2001. [Greedy function approximation: A gradient boosting machine](#). *Ann. Statist.*, 29(5):1189–1232.
- Ben He and Iadh Ounis. 2006. [Query performance prediction](#). *Information Systems*, 31(7):585 – 594. (1) SPIRE 2004 (2) Multimedia Databases.
- Alex Hern. 2017. [Facebook translates 'good morning' into 'attack them', leading to arrest](#).
- Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng Li. 2019. Read+ verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6529–6537.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *EMNLP*.
- Robin Jia and Wanze Xie. 2020. Know when to abstain calibrating question answering system under domain shift. Technical report, Stanford University.
- Jin Huang and C. X. Ling. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310.
- Rebecca Klar. 2019. [Google under fire for mistranslating chinese amid hong kong protests](#).
- Bernhard Kratzwald, Anna Eigenmann, and Stefan Feuerriegel. 2019. Rankqa: Neural question answering with answer re-ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6076–6085.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural Questions: a benchmark for question answering research](#). *TACL*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. [Albert: A lite bert for self-supervised learning of language representations](#).
- Tong Niu and Mohit Bansal. 2019. [Automatically learning data augmentation policies for dialogue tasks](#).
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.
- Lin Pan, Rishav Chakravarti, Anthony Ferritto, Michael Glass, Alfio Gliozzo, Salim Roukos, Radu Florian, and Avirup Sil. 2019. [Frustratingly easy natural question answering](#).
- Anselmo Peñas, Álvaro Rodrigo, and Felisa Verdejo. 2007. Overview of the answer validation exercise 2007. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 237–248. Springer.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). *EMNLP*.
- Chuanqi Tan, Furu Wei, Qingyu Zhou, Nan Yang, Weifeng Lv, and Ming Zhou. 2018. I know there is no answer: modeling answer validation for machine reading comprehension. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 85–97. Springer.
- Ming Ning Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *EMNLP/IJCNLP*.
- Yongquan Tao and Shengli Wu. 2014. [Query performance prediction by considering score magnitude and variance together](#). In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14*, page 1891–1894, New York, NY, USA. Association for Computing Machinery.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. [Universal adversarial triggers for attacking and analyzing NLP](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2017. Evidence aggregation for answer re-ranking in open-domain question answering. *ICLR*.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. *NAACL*.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, and Hai Zhao. 2020a. Sg-net: Syntax-guided machine reading comprehension. *AAAI*.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020b. Retrospective reader for machine reading comprehension. *arXiv preprint arXiv:2001.09694*.