Using contradictions to improve QA systems

Anonymous EMNLP submission

Abstract

Ensuring the safety of question answering (QA) systems is critical for deploying them in biomedical and scientific domains. One approach to improving these systems uses natural language inference (NLI) to determine whether answers are supported, or entailed, by some background context. However, these systems are vulnerable to supporting an answer with a source that is wrong or misleading. Our work proposes a critical approach by selecting answers based on whether they have been contradicted by some background context. We evaluate this system on multiple choice and extractive QA and find that while the contradiction-based systems are competitive with and often better than entailment-only systems, models that incorporate contradiction, entailment, and QA model confidence scores together are the best. Based on this result, we explore unique opportunities for leveraging contradiction-based approaches such for improving interpretability and selecting better answers.

1 Introduction

Safety in NLP systems is an unresolved issue, particularly in biomedical and scientific contexts where known issues such as hallucination and overconfidence provide obstacles for deploying them (Ji et al., 2022; Kell et al., 2021). Utilizing natural language inference (NLI) as a method for improving the safety and performance of NLP research is an active area of research (Li et al., 2022). However, these systems typically focus exclusively on entailment to verify answers. Similar research looks at building self-supporting NLP systems (Nakano et al., 2022; Menick et al., 2022) with the goal of improving safety by verifying model outputs with some external supporting source.

These developments are troubling since a verification or self-supporting approach is vulnerable to selecting supporting sources that might be wrong.

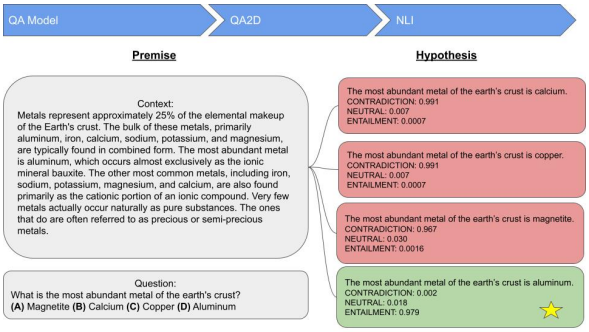

Figure 1: A QA model is used to produce answers which are reformulated as hypothesis statements to determine if they are entailed or contradicted by a premise. The answers are ranked by the NLI class scores to select the best answer.

Since supported or verified answers look more credible, a user might be mislead into uncritically accepting model outputs then they otherwise would be. Even though we could also find sources that wrongly contradict an answer, surfacing sources that contradict an answer might help a user engage critically and help a QA system select the least contradicted answers. Therefore we ask: under NLI-based setups how do contradictions contribute to the performance of question answering (QA) and how is this different from entailment-based systems? By exploring this question, we hope to show why researchers should be critical of the paradigm of verification in NLP systems and why future work utilizing critical and contradicted statements could provide unexplored opportunities for improving NLP systems.

Our work makes the following contributions. We propose a method (Figure 1) that reformulates answers under QA as hypothesis statements which are then used with three-class NLI to rank and select answers. We demonstrate, across 9 multiple choice datasets and 7 extractive QA datasets models, that models which use QA confidence scores as well as both entailment and contradiction scores

outperform all other setups. In addition, selecting the least contradicted answer provides a competitive approach to selecting answers that is often on par with or better than entailment-based systems. While this work is in a relatively limited setting, we suggest how leveraging contradictions could help improve QA inference in ways that are not possible with entailment-based systems.

1.1 Related work

NLI for QA has been explored by several authors (see the overview in Paramasivam and Nirmala (2021)) showing performance improvements in multiple choice (Mishra et al., 2021), extractive (Chen et al., 2021), open domain (Harabagiu and Hickl, 2006) and multihop (Trivedi et al., 2019) settings. These approaches have thus far focused on using entailment as a verification mechanism. Chen et al. (2021) find that under selective question answering (Kamath et al., 2020) for extractive QA, NLI systems can verify QA systems' predictions. However, their result is limited to only selecting a top k % of answers and they do not provide an approach for improving QA systems overall performance nor show what their results would have been like if they incorporated the contradiction signal. Mishra et al. (2021) explores the use of entailment for multiple choice and fact checking settings and finds that not only do NLI models do well at these tasks by themselves but when they are adapted using in-domain data and longer premises they perform even better. Despite this, Mishra et al. (2021) uses a two-class NLI set up (entailed or not entailed) which means there would be no information about the effect of using the contradiction class if this approach was used.

Factual consistency is the only domain that leverages contradictions directly. Factual consistency seeks to ensure that a collection of utterances do not contain contradictions such as unfaithfulness towards a source document (see Li et al. (2022) for an overview). Here approaches to improve faithfulness are still focused on entailment. Laban et al. (2022) proposes a NLI-based method to ensure the consistency of a summary with a source document that incorporates contradiction and neutral scores with entailment scores beating out previous systems. Interestingly, they show that a combination of entailment and contradiction achieves the best results over entailment alone. Similarly, QAFactEval (Fabbri et al., 2022) improves on Laban et al. (2022) and maintains the approach of incorporating all NLI class scores. Schuster et al. (2022) and Hsu et al. (2021) develop interesting cases where contradictions are leveraged to identify consistency errors within or across wikipedia articles illustrating the further utility of contradictions. Finally, contradiction detection has surfaced as an important tool in generating dialogues that are consistent with a persona (Nie et al., 2021; Song et al., 2020). To our knowledge, this is the first work to directly leverage contradictions for QA.

2 Method

2.1 Overview

The proposed approach is similar to Chen et al. (2021) and Mishra et al. (2021) where question answer pairs are turned into declarative statements (QA2D) (see Demszky et al. (2018)). QA models for each setting are used to produce answers and confidence scores for each answer which are later used to train calibration models. Three-class NLI classification (entailment, neutral, contradiction) is performed on the provided QA contexts (the premises) with the hypotheses produced from the earlier QA2D model. Like Chen et al. (2021) a calibration method is used that combines the confidence scores from the QA and NLI models. In the multiple choice setting, answers are selected among a set of alternatives for a given question through ranking by a score produced by the models above. In the extractive QA case, questions are ranked for selective QA (Kamath et al., 2020) where a top k number of answers pre-selected by a QA model are selected by how confident the model is in answering a question.

2.2 QA models

For the multiple choice setting, we used RoBERTa large (Liu et al., 2019) finetuned on the RACE dataset (Lai et al., 2017) as well as two DeBERTa v3 (He et al., 2021a) variants (base and xsmall) finetuned on SciQ (Welbl et al., 2017). For the extractive QA setting, we used DistillBERT (Sanh et al., 2020) and BERT-Large (Devlin et al., 2019)) models trained on SQuAD (Rajpurkar et al., 2016). More details on model selection, training, and validation are available in Appendix A. In both cases, answers are selected given a context provided by the dataset and those contexts are used as the premise for NLI.

2.3 QA2D

A QA2D model reformulates a question-answer pair to a declarative statement (Demszky et al., 2018). As noted in Chen et al. (2021) and Mishra et al. (2021), the QA2D reformulation is critical to using NLI models in QA since the proposed answer needs to match the format of NLI. We trained a T5-small model (Raffel et al., 2020) on the dataset proposed by Demszky et al. (2018) for QA2D since we found almost no noticeable differences in performance in larger models. Unlike Chen et al. (2021), we found that regardless of size, these QA2D models struggled with long questions or questions with complex syntax and would often leave the answer out of the statement. In order to solve this, constrained decoding that required the answer to be in the statement was tried. However, this often produced ungrammatical or nonsensical statements. We settled with the following heuristic to postprocess QA2D outputs: If less than 50% of the tokens in the answer were in the statement then we appended the answer to the end of the statement. 50% was used to account for rephrasing the answer or swapping pronouns. While some statements resulted in answer redundancy, this was better than having hypotheses which left out the answer. Future work on QA2D should focus on how these models can be used outside of the domains in the dataset provided by Demszky et al. (2018).

2.4 NLI

NLI is then used to classify whether the reformulated answer is contradicted, entailed, or neutral w.r.t to a context passage. The whole context was used as Schuster et al. (2022) and Mishra et al. (2021) demonstrated that long premises still performed adequate though not as well as sentence-length premises. Using the whole context avoids needing to use decontextualization as is required in Chen et al. (2021). We used two DeBERTa-based models (He et al., 2021b) trained on the MNLI dataset (Williams et al., 2018) (called mnli-base and mnli-large) and an ALBERT model (Lan et al., 2019) trained on the ANLI dataset in addition to various other NLI datasets (called albert-anli) (Nie et al., 2020). After inference, the confidence scores are then used for each class in the procedures below.

2.5 Calibration models

Like Kamath et al. (2020) and Chen et al. (2021) we developed a set of calibration models in order to do answer ranking. A calibration model is trained on a set of posterior probabilities from downstream models to predict whether an answer is correct. To compare the effect of using different combinations of NLI class confidence scores we trained a logistic regression model on linear combinations of the following features: **QA** indicates that the QA model confidence score is being used, **E** indicates the entailment score is used, **C** indicates the contradiction score is used and **N** indicates the neutral score. Like Chen et al. (2021), all calibration models are trained on a holdout set of 100 samples from a single domain using logistic regression which predicts, given the confidence scores of the downstream models, whether the answer is correct. A multi-domain calibration approach like in Kamath et al. (2020) was not used since the focus was a minimum experiment to test the viability of leveraging different NLI classifications. To illustrate the characteristics of the calibration models, Appendix D presents a regression analysis for the multiple choice setting.

2.6 Answer ranking

Similar to Harabagiu and Hickl (2006), answers are ranked based on the highest probability from the calibration model σ given a linear combination of the QA or NLI scores given an answer $n \in N$ answer set. When a single feature is used such as NLI class or QA class no calibration is made and σ is simply the identity of the confidence score. In the case of contradiction only σ is the inverse of the contradiction confidence score, indicating the least contradicted answer is being selected. Formally our approach can be described as:

$$\underset{N}{\operatorname{argmax}}\, \sigma(QA_n; NLI_n)$$

For the multiple choice setting we used this for selecting the answer to a given question among a set of alternative answers N. We found that using a top $K = 4$ approach to extractive QA produced almost the same answer alternatives with slightly different spans so we did not use the alternatives ranking approach with extractive QA. For both the multiple choice and extractive QA settings we ranked answers like in Kamath et al. (2020), where a top n set of questions at a certain coverage coverage

threshold k is selected, resulting in a set of top answers the model is most confident in answering.

2.7 Datasets

For both settings, datasets where the context passage is already available were used. For the multiple choice setting a set of 9 datasets were used. Two of those datasets are in-domain for the QA and calibration, RACE and SciQ. For extractive QA 5 of the datasets from the MRQA 2019 task were selected (Fisch et al., 2019) as well as SQuAD 2.0 (Rajpurkar et al., 2018) and SQuAD adversarial (Jia and Liang, 2017) for a total of 7 extractive QA datasets. The in-domain dataset for the extractive QA model is SQuAD and the dataset used for calibration is Natural Questions since that is what was used by Chen et al. (2021). The only preprocessing done was to remove questions where the context was empty. Appendix B describes full details on the datasets used for evaluation.

3 Results

3.1 Ranking multiple choice

In tables 1 (NLI only) and 2 (calibrated) we present the accuracy achieved on each of the 9 datasets for the mutiple choice setting. We show results with each QA model and the mnli-large model for NLI (Appendix C shows results for alberta-anli and mnli-base which perform worse but generally reflect the same trends). On the NLI-only results presented in table 1, the RoBERTA-RACE QA model outperforms other approaches on most datasets except MCTest and the combination of entailment and contradiction tends to do second best. Notably, the SciQ models do much worse than the NLI-only ranking for either class except on in-domain questions for SciQ and the similar QASC dataset. This could possible a result of the RACE domain being more generic than the SciQ domain or the RACE dataset being larger. The results show that an NLI-only approach can be competitive with a robust QA model and better than a more limited QA model. Notably, incorporating the contradiction scores with the entailment scores is better than entailment alone and that selecting the least contradicted answer is quite competitive with selecting the most entailed answer.

The results from the calibration models show that the NLI calibrated models outperform QA only in all cases (they perform the same as RoBERTA-RACE in the in-domain case). The best calibration incorporates QA confidence, entailment, and contradiction (QA+E+C) achieving an average accuracy of 84.57% over 84.09% from RoBERTa-RACE. The second best approach is the calibration with contradiction only (84.33%), however only slightly over entailment only (84.31%). To inspect these trends further, a correlation analysis in Appendix E is provided on how each NLI class and QA confidence score correlates with the correct answer. Interestingly, other than QA model confidence scores, it is contradiction confidence score that has the strongest correlation with the correct answer further demonstrating the utility of leveraging contradictions.

3.2 Selective QA

For selective QA evaluation in both settings, the QA model selects the answer and then we evaluate the top 20% or 50% of those answers after sorting them by the approaches we outlined earlier. In the multiple choice setting we do not select the answers for individual questions with the ranking as proposed in the method above since the approach under performed the QA model. Those results are available in Appendix C.

3.2.1 Selective QA for multiple choice QA

In table 3, the first thing to note is that the QA + C model performs the best on selective QA, achieving best or second best accuracy on almost every dataset for an average of 97.57% at 20% coverage and 94.74% at 50% coverage over 97.45% at 20% coverage and 94.52% at 50% coverage achieved by the RoBERTA-RACE QA model. This is especially striking at 50% coverage where the QA model only does significantly better on the in-domain RACE datasets. The QA+C model is the only model to outperform the QA model ranking by confidence score. The NLI only models can be competitive with the calibrated models. Sorting by the least contradicted achieves good performance and is often better (94.79% @ 20% / 92.13% @ 50%), than sorting by the most entailed (93.50% @ 20% / 91.24% @ 50%) or a combination of entailed and contradicted (93.55% @ 20% / 91.89% @ 50%). These results are inline with our intuition that the less contradicted an answer is the more likely it is correct even in cases where there is uncertainty about its entailment.

Model	Cosmos	DREAM	MCS	MCS2	MCT	QASC	RACE	R_C	SciQ	Avg
s-base	18.46	43.80	61.99	63.71	44.76	93.41	30.97	27.39	95.28	53.30
s-small	25.46	48.26	60.28	66.04	59.76	90.60	35.56	30.62	98.09	57.18
QA	**64.22**	**82.56**	**89.70**	**86.98**	90.48	**98.16**	**76.93**	**69.80**	**97.96**	**84.08**
E+C	44.36	80.94	85.52	84.99	**90.60**	96.44	64.29	51.40	92.47	76.77
E	36.18	79.03	86.02	79.72	89.88	95.90	62.14	49.72	91.96	74.50
C	59.26	78.98	83.12	84.43	89.29	92.76	62.74	47.05	91.58	76.58

Table 1: Accuracy scores on NLI-only answer ranking. RoBERTa-RACE is indicated as QA. The best scores are bold and the second best are underlined.

Model	Cosmos	DREAM	MCS	MCS2	MCT	QASC	RACE	R_C	SciQ	Avg
s-base	18.46	43.80	61.99	63.71	44.76	93.41	30.97	27.39	95.28	53.30
s-small	25.46	48.26	60.28	66.04	59.76	90.60	35.56	30.62	98.09	57.18
QA	64.22	82.56	89.70	86.98	90.48	98.16	76.93	**69.80**	97.96	84.08
QA+E+C	**64.72**	**83.19**	**90.06**	**87.59**	**91.43**	**98.60**	**77.53**	**69.80**	**98.21**	**84.57**
QA+E	64.32	82.85	89.92	87.29	91.07	98.49	77.18	69.66	98.09	84.31
QA+C	64.82	82.75	89.88	87.29	90.83	98.38	77.16	**69.80**	98.09	84.33

Table 2: Accuracy scores on calibrated NLI answer ranking. Calibrations are with the RoBERTa-RACE model (QA). The best scores are bold and the second best are underlined.

	Database	QA +E+C	QA+C	QA+E	E+C	E	C	QA
20%	CosmosQA	77.55	**91.12**	76.88	69.18	68.34	83.25	88.61
	DREAM	98.28	**98.77**	98.28	96.32	96.32	96.81	98.28
	MCScript	**99.82**	99.46	**99.82**	99.64	99.64	99.46	**99.82**
	MCScript-2.0	99.58	**99.72**	99.45	99.17	99.03	97.37	99.58
	MCTest	**100**	99.40	**100**	**100**	**100**	99.40	98.81
	QASC	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	RACE	94.93	96.69	94.72	92.44	92.24	90.17	**98.24**
	R_C	88.73	92.96	89.44	85.21	85.92	86.62	**93.66**
	SciQ	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	Avg	95.43	**97.57**	95.40	93.55	93.50	94.79	97.45
50%	CosmosQA	80.29	**81.70**	76.94	75.80	70.64	80.63	76.47
	DREAM	95.10	**96.86**	94.90	93.63	93.63	93.63	96.67
	MCScript	98.57	98.64	98.28	98.00	97.93	97.14	**98.78**
	MCScript-2.0	96.40	**98.23**	95.84	94.68	94.40	96.01	98.01
	MCTest	99.52	**99.76**	99.52	99.05	99.05	99.76	99.52
	QASC	**100**	**100**	**100**	99.78	99.78	99.78	**100**
	RACE	90.11	92.68	89.99	87.71	87.38	85.23	**93.88**
	R_C	85.11	84.83	85.39	78.37	78.37	77.25	**87.36**
	SciQ	**100**	**100**	**100**	**100**	**100**	99.74	**100**
	Avg	93.90	**94.74**	93.43	91.89	91.24	92.13	94.52

Table 3: Selective QA for the multiple choice with accuracy scores at 20% and 50% coverage of the dataset. Calibrations and QA confidence are all from RoBERTa-RACE where RACE is the in-domain dataset.

	Dataset	QA+E+C	QA+E	QA+C	E+C	E	C	QA
20%	BioASQ	85.04	**85.06**	83.10	74.22	74.22	75.47	82.99
	HotpotQA	86.62	**86.69**	85.89	80.60	80.60	79.82	85.33
	NaturalQuestions	91.84	91.68	**92.18**	79.89	79.87	82.09	90.98
	SQuAD	98.26	98.76	98.17	92.37	92.48	90.88	**99.04**
	SQuAD-adv	**43.99**	43.98	43.57	43.74	43.60	42.81	39.83
	SQuAD2	37.64	37.56	36.07	37.43	37.31	**37.68**	30.52
	TriviaQA	**81.33**	81.21	80.36	65.53	65.25	69.13	80.68
	Avg	74.96	**74.99**	74.19	67.68	67.62	68.27	72.77
50%	BioASQ	**76.13**	76.04	75.51	71.49	71.49	72.97	75.49
	HotpotQA	**79.37**	79.30	78.95	77.43	77.43	77.31	78.74
	NaturalQuestions	**84.53**	84.48	83.24	74.96	74.93	78.62	82.47
	SQuAD	96.98	96.97	97.01	91.58	91.52	91.19	**97.00**
	SQuAD-adv	41.80	41.16	41.49	42.76	**42.79**	42.03	40.26
	SQuAD2	29.41	28.45	28.77	**34.43**	34.14	34.39	26.18
	TriviaQA	74.30	**74.37**	74.23	65.05	64.93	68.08	74.21
	Avg	**68.93**	68.68	68.46	65.39	65.32	66.37	67.76

Table 4: Selective QA for extractive QA with F1 scores at 20% and 50% coverage. Calibrated models and QA use the BERT-large model.

3.2.2 SelectiveQA for extractive QA

For the extractive QA setting we present the same analysis in table 4. Similar trends to multiple choice QA are present where calibration with contradiction only, QA + C, has better average F1 scores than the QA model (74.19% vs 72.77% at 20%, 68.46% vs 67.76% at 50%). Of the NLI-only ranking, selecting the least contradicted does best. Although only slightly better than the QA + C, the results show that the QA + E model does best at 20% coverage and the QA + E + C model does best at 50% coverage. This indicates that entailment is still an important signal, albeit more powerful when combined with contradiction. Appendix C contains a comparison with the smaller DistillBERT model which shows similar results. Notably with a smaller model QA+E+C does best in all cases and that selecting the least contradicted answer without any calibration does second best at 50% coverage.

3.3 Answer Rejection on SQuAD 2.0

In order to explore how useful contradiction might be in other settings, we evaluated the answer rejection task in SQuAD 2.0 (Rajpurkar et al., 2018) using our BERT-large model. This task evaluates how well a model does at abstaining from answering a question that is unanswerable. Three setups are used: rejecting answers by QA confidence, by entailment score, and by contradiction score. When selecting by least entailed answers the problem becomes a two-class NLI (entailed v not entailed) which was previously looked at by Chen et al. (2021).

Table 5 shows that the NLI-based setups outperform QA confidence setups in all cases. Interestingly, the difference between rejecting answers that are not entailed and rejecting answers that have been contradicted appears to reflect a precision versus recall trade off. The overall best model (best F1 score) is achieved by rejecting answers where the contradiction score was greater than 5%, successfully rejecting 76.15% answers and accepting 93.23% answerable questions. Rejecting answers if they are not entailed, where E < 50%, achieves the second best F1 score and illustrates an interesting dynamic. E < 50% has the best recall (38.52%), successfully rejecting the most answers, while C > 50% has the best precision (97.06%), accepting the most answerable questions. This result shows that if we want to build systems that err on the side of rejecting answers not entailed has an advantage. Conversely, if we want to build systems that are better at rejecting only answers that should be rejected then contradicted is a better strategy. The results highlight the utility of using contradiction confidence scores even if they are low which gives credence to using the contradiction score as a meaningful signal.

	Reject	Accept	Precision	Recall	F1
QA <50%	46.71%	86.15%	62.81%	23.39%	34.09%
QA <25%	22.29%	95.45%	71.05%	11.16%	19.29%
QA <75%	71.22%	72.86%	56.79%	35.66%	43.81%
E <5%	43.80%	98.74%	94.55%	21.93%	35.61%
E <25%	63.82%	96.58%	90.33%	31.95%	47.21%
E <10%	52.14%	98.02%	92.95%	26.11%	40.77%
E <50%	**76.94%**	91.52%	81.96%	**38.52%**	52.41%
C >50%	42.78%	**99.35%**	**97.06%**	21.42%	35.09%
C >25%	54.21%	98.59%	95.05%	27.15%	42.23%
C >10%	66.88%	96.50%	90.53%	33.49%	48.89%
C >5%	76.15%	93.23%	84.92%	38.13%	**52.63%**

Table 5: Rejecting unanswerable questions in SQuAD2.0 (11,873 answers total with 5,945 unanswerable questions). Bold indicates the best score and underlined indicates the second best score.

4 Discussion

While the results above show that contradiction is an important signal for improving performance of QA systems in the settings above, contradiction provides additional unique opportunities for improving NLP systems overall. Contradiction is a particularly important signal because it can improve interpretability. When choosing answers based on the least contradicted answer, we have information about the other answers and why we didn't select them. Namely, that they were contradicted. Entailment and QA model confidence do not have the same interpretability since all that is known about the other answers is they have a lower entailment or confidence score, they could still be correct or entailed. In addition, we would not know if the unselected alternatives were neutral or contradicted with respect to the premise.

Once an answer is known to be contradicted, that information can be used to try retrieving another answer. In models that support prompting, we can use that contradiction as a hint for another attempt at an answer. Entailment does not lend itself to this iterative refinement of question answering and we suggest that future work on utilizing contradiction should investigate developing inference techniques that take advantage of the contradiction signal.

Contradiction also provides a unique opportunity for open domain QA systems which require retrieving a context containing the answer. Like entailment-based approaches (Harabagiu and Hickl, 2006) we can try selecting the least contradicted passage for a downstream reader. We can also imagine extending the work of Schuster et al. (2022) where contradiction-based approaches could be used to retrieve passages that would contradict an answer to determine if the proposed answer might be wrong and thereby develop an iterative inference procedure for open domain settings. Retrieved contradicting sources could also be surfaced to a user to help them critically engage with selected answers by the model.

Finally contradicted statements are already being used in a generative setting to improve fact verification systems during train time (Wright et al., 2022; Pan et al., 2021; Saakyan et al., 2021). Recent work (Saunders et al., 2022) has shown that self-criticism is a powerful technique for improving the quality of NLP systems during inference and we believe generating critical statements for model predictions could help with overall performance, interpretability, and safety by providing outputs with a full picture under which they might be faulty. Future work should assess whether a contradiction-based approaches to improve NLP safety along these lines is an interesting alternative to the current verification-based approaches.

5 Limitations

Despite the results above, multiple choice QA and extractive QA with a provided context is a limited setting that doesn't indicate the results would extend to other popular settings where NLI. Given that Laban et al. (2022) shows similar results that contradiction is an important signal in factual consistency we are hopeful that it would.

5.1 Context Length and NLI datasets

Even though there is a greater tendency to use NLI in zero-shot settings (Yin et al., 2020). Domain

transfer is a known issue with using NLI models. In particular, NLI datasets tend to focus on textual entailment over short passages such as sentence pairs and performance degrades when using longer passages such as in the datasets we use (Mishra et al., 2021). Even when in-domain datasets are created (Chen et al., 2021; Khot et al., 2018; Mishra et al., 2021). They tend to focus on data augmentation strategies that produce two-class NLI datasets (entail, not entailed) which wouldn't give us any contradiction signals. Future work should pick up on producing models capable of performing textual entailment over longer passages and devising methods for generating three-class NLI datasets so that we can determine if contradictions receive the same benefits from those approaches that entailment has. Additionally we saw that albert-anli performed worse than mnli-large and mnli-large performs poorly on some datasets indicating that we still have much more work to do to improve upon NLI in general.

5.2 Ranking requires alternatives and time

In the extractive QA setting presented above we did not use ranking answer alternatives like we used for the multiple choice setting due to lack of more diverse outputs. Further work with extractive QA models that produce diverse alternatives is required. Like other textual entailment based systems, this speaks to the computational expense involved in generating and evaluating answer alternatives. If we were to apply our method to an open domain setting where a set of context passages are retrieved, the ranking procedure would require a quadratic evaluation procedure for each context passage against each reformulated answer candidate (Schuster et al., 2022). Future work should look towards comparison approaches that amortize the computational cost involved in pairwise NLI-based ranking techniques such as investigating NLI-based dense passage retrieval (Reimers and Gurevych, 2019).

6 Summary

We have demonstrated that incorporating contradiction is an important signal for multiple choice and extractive QA systems. By proposing a method that reformulates answers as hypothesis statements, the system is able to rank answers and demonstrate that QA model confidence calibrated with entailment and contradiction scores outperform QA models by themselves as a ranking approach in all cases. In addition, models calibrated with contradiction only or simply selecting the least contradicted answers with NLI only provides a competitive approach to selecting answers that is often on par with or better than entailment-only systems. These results show that we should rethink the paradigm of verifying answers with entailment across NLP setups. While this work is in a relatively limited setting, we provide discussion on how leveraging contradictions could help improve open domain QA as well as other NLP tasks at large.

7 Ethics Statement

Works addressing NLP safety should be aware of their limitations and be clear about potential harms and misuse of their proposed approaches. Systems that improve safety by verification or support are vulnerable to drawing on untrue and biased sources to justify their outputs. The appearance of credibility given to texts that use citations and appeals to authority means that users should be made aware that the sources they draw on can be wrong. This applies to critical sources as well since a source can provide criticism that is wrong or misleading. However, by presenting contradictions we believe that systems could provide a wider breadth of options for users to engage with critically than models which claim to verify answers by appealing to the authority of a source document.

References

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI Models Verify QA Systems' Predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming Question Answering Datasets Into Natural Language Inference Datasets. Technical Report arXiv:1809.02922, arXiv. ArXiv:1809.02922 [cs] type: article.

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.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization. Number: arXiv:2112.08542 arXiv:2112.08542 [cs].

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for Using Textual Entailment in Open-Domain Question Answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. Number: arXiv:2111.09543 arXiv:2111.09543 [cs].

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. Number: arXiv:2006.03654 arXiv:2006.03654 [cs].

Cheng Hsu, Cheng-Te Li, Diego Saez-Trumper, and Yi-Zhan Hsu. 2021. WikiContradiction: Detecting Self-Contradiction Articles on Wikipedia. Technical Report arXiv:2111.08543, arXiv. ArXiv:2111.08543 [cs] type: article.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. Number: arXiv:2202.03629 arXiv:2202.03629 [cs].

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.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective Question Answering under Domain Shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Gregory Kell, Iain Marshall, Byron Wallace, and Andre Jaun. 2021. What Would it Take to get Biomedical QA Systems into Practice? In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 28–41, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A Dataset for Question Answering via Sentence Composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8082–8090. Number: 05.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In *AAAI*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods. Number: arXiv:2203.05227 arXiv:2203.05227 [cs].

Yichan Liang, Jianheng Li, and Jian Yin. 2019. A New Multi-choice Reading Comprehension Dataset for Curriculum Learning. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, pages 742–757. PMLR. ISSN: 2640-3498.

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. Number: arXiv:1907.11692 arXiv:1907.11692 [cs].

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers

with verified quotes. Number: arXiv:2203.11147 arXiv:2203.11147 [cs].

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking Beyond Sentence-Level Natural Language Inference for Question Answering and Text Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. Number: arXiv:2112.09332 arXiv:2112.09332 [cs].

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A Machine Comprehension Corpus Focused on Script Events and Participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot Fact Verification by Claim Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

Aarthi Paramasivam and S. Jaya Nirmala. 2021. A survey on textual entailment based question answering. *Journal of King Saud University - Computer and Information Sciences*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Number: arXiv:1910.01108 arXiv:1910.01108 [cs].

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. Number: arXiv:2206.05802 arXiv:2206.05802 [cs].

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching Sentence-pair NLI Models to Reason over Long Documents and Clusters. Number: arXiv:2204.07447 arXiv:2204.07447 [cs].

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating Persona Consistent Dialogues by Exploiting Natural Language Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8878–8885. Number: 05.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231. Place: Cambridge, MA Publisher: MIT Press.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing Entailment for Multi-Hop Question Answering Tasks. 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 2948–2958, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *NUT@EMNLP*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. Number: arXiv:1910.03771 arXiv:1910.03771 [cs].

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating Scientific Claims for Zero-Shot Scientific Fact Checking. Technical Report arXiv:2203.12990, arXiv. ArXiv:2203.12990 [cs] type: article.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online. Association for Computational Linguistics.

A Training Setup and Reproducibility

Table 7 outlines the pretrained models that we used and datasets they are trained on, all of these models are publicly available on the huggingface model hub under the locations listed. Where space doesn't allow RoBERTa-RACE is aliased as RACE. In addition to several pretrained models used in the setups described earlier, we trained 3 models, a t5-small model on Demszky et al. (2018) for the QA2D set up where a Rogue1 of 90.73% is achieved on the validation set, DeBERTa-v3 models (xsmall and base) trained on SciQ (Welbl et al., 2017) achieving 93.99% accuracy on the xsmall model and 91.76% accuracy on the base model. Where space doesn't allow the DeBERTa-v3 models are called s-base and s-small. All models were trained using the huggingface trainer API (Wolf et al., 2020) with an Adam optimizer at a learning rate of 5.60e-05 with weight decay of 0.01. All models and inference were performed on 1 Tesla P100 GPU. More details are available on table 7. Full instructions on reproducibility as well as trained models are provided in the publicly available code including directions to weights and biases to inspect the training runs, full parameter set, and evaluations suites which will be available upon publication.

B Dataset Details

The tables (Table 8 and Table 9) below outline the datasets used . Additional details such as train size and preprocessing steps are available in the references provided. When space doesn't allow CosmosQA is aliased to Cosmos, MCScript to MCS, MCScript-2.0 to MCS2, MCTest to MCT, and RACE-C to R_C. As mentioned previously the only preprocessing step used was to filter out questions where no context passage is provided. Finally, validation splits are used in the CosmosQA and QASC case since context passages or gold answers are not made available so readers should be aware of this when reading results on those datasets.

C Model size and approach performance analysis

In order to help understand how the results presented above differ with model size or approach

Huggingface	Name
LIAMF-USP/roberta-large-finetuned-RACE	RoBERTa-RACE
bert-large-uncased-whole-word-masking-finetuned-squad	BERT-Large
distilbert-base-uncased-distilled-squad	DistillBERT
ynie/albert-xxlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli	albert-anli
microsoft/deberta-base-mnli	mnli-base
microsoft/deberta-v2-xxlarge-mnli	mnli-large

Table 6: Pretrained models that we used.

Model	Dataset	Epochs	Score	
t5-small	Demszky et al. (2018)	20	Rogue1	90.73
deberta-v3-xsmall	Welbl et al. (2017)	6	Accuracy	93.99
deberta-v3-base	Welbl et al. (2017)	6	Accuracy	91.79

Table 7: The models we trained for or setups with evaluation scores and number of epochs trained.

Dataset	Split	Size	Reference
CosmosQA	validation	2985	Huang et al. (2019)
DREAM	test	2041	Sun et al. (2019)
MCScript	test	2797	Ostermann et al. (2018)
MCScript-2.0	test	3610	Ostermann et al. (2019))
MCTest	test	840	Richardson et al. (2013)
QASC	validation	926	Khot et al. (2020)
RACE	test	4934	Lai et al. (2017)
RACE-C	test	712	Liang et al. (2019))
SciQ	test	884	Welbl et al. (2017)

Table 8: Datasets used for the multiple choice setting including split used and sample size. Validation splits were used in the case of CosmosQA since the test split is not publicly available and QASC since context passages or gold answers are not available.

we have presented the following supplemental tables. Table 10 shows differences in performance between mnli-base, mnli-large, and albert-anli. Table 11 shows selective QA accuracies in the multiple choice setting where answer selection is done through ranking before we rank answers for selective QA. Selective QA on extractive QA using DistillBERT (table 12) shows that QA+E+C does best in all cases and contradiction only does second best at 50% coverage.

D Regression Analysis

Table 13 shows a supplemental regression analysis for each calibration model used in the multiple choice settings. The results indicate that as the MNLI model gets larger the calibration model uses its NLI confidence scores more. Importantly entailment coefficients are stronger than contradiction coefficients in all cases and this should be kept in mind when considering the results presented in this paper.

E Correlation Analysis

Since we are using the NLI and QA model scores to construct the setups above, we'd like to know how these factors correlate with the correct answer. Table 15 shows how each NLI class correlates both by score and by actual classification (score > 50%) as compared against QA model confidence score. The multiple choice analysis shows answers from the RoBERTa-RACE model and the extractive QA analysis shows answers from the BERT-large model trained on SQuAD. The correlation analysis presents Spearman rank correlations. What we see is that in the multiple choice setting the confidence score has a strong correlation with the correct answer which makes sense given the confidence score is a softmax over the multiple choice classes. Extractive QA confidence scores have a much weaker correlation and tend to have

Dataset	Size	Reference
BioASQ	1504	Fisch et al. (2019)
TriviaQA	7785	
HotpotQA	5901	
SQuAD	10506	
Natural Questions	12836	
SQuAD2	11871	Rajpurkar et al. (2018)
SQuAD-adv	5347	(Jia and Liang, 2017)

Table 9: Extractive QA datasets used. Validation sets are used on the SQuAD2.0 and SQuAD adversarial datasets and MRQA 2019 dev sets are used for the MRQA 2019 sets.

Model	Cosmos	Dream	MCS	MCS2	MCT	QASC	RACE	R_C	SciQ	Avg
SciQ-base	18.46	43.80	61.99	63.71	44.76	93.41	30.97	27.39	95.28	53.31
SciQ-small	25.46	48.26	60.28	66.04	59.76	90.60	35.56	30.62	98.09	57.19
RACE	64.22	82.56	89.70	86.98	90.48	98.16	76.93	69.80	97.96	84.09
mnli-base										
QA + E + C	64.32	82.66	89.63	87.01	90.71	98.27	76.95	69.80	98.09	84.16
QA + E	64.25	82.66	89.63	86.98	90.71	98.27	76.95	69.80	97.96	84.14
QA + C	64.29	82.56	89.63	87.01	90.60	98.16	76.93	69.80	97.96	84.1
E + C	33.03	62.27	76.76	72.11	68.57	92.66	45.16	34.41	88.01	63.66
E	27.81	62.47	79.37	71.94	68.81	92.66	43.48	34.41	88.01	63.22
C	43.45	59.19	70.18	69.97	67.50	81.86	41.81	32.58	87.37	61.55
albert-anli										
QA + E + C	64.19	82.56	89.70	87.06	90.48	98.16	76.93	69.80	97.96	84.09
QA + E	64.19	82.56	89.70	87.06	90.60	98.16	76.93	69.80	97.96	84.11
QA + C	64.22	82.56	89.70	86.98	90.48	98.16	76.93	69.80	97.96	84.09
E + C	35.71	68.20	79.55	73.88	77.50	91.79	49.05	39.47	90.82	67.33
E	33.67	68.35	79.91	73.19	77.38	91.90	49.07	39.19	90.94	67.07
C	45.16	63.74	73.58	72.71	73.33	77.86	46.34	38.20	87.24	64.24

Table 10: Accuracy scores in the multiple choice setting for various NLI models used. Calibration was with the RoBERTA-RACE model.

	Dataset	QA+E+C	QA+E	QA+C	E+C	E	C	QA
20%	CosmosQA	77.55	67.17	83.25	20.10	27.47	67.50	**88.61**
	DREAM	**98.28**	96.32	96.81	81.13	91.91	93.87	**98.28**
	MCScript	**99.82**	**99.64**	99.46	93.02	98.93	96.96	**99.82**
	MCScript-2.0	**99.58**	99.03	97.37	92.24	97.37	95.01	**99.58**
	MCTest	**100**	**100**	99.40	85.12	97.02	97.02	98.81
	QASC	**100**	**100**	**100**	97.30	**100**	99.46	**100**
	RACE	94.93	92.13	90.17	62.73	76.71	75.05	**98.24**
	RACE-C	88.73	85.21	86.62	71.13	74.65	69.01	**93.66**
	SciQ	**100**	**100**	**100**	82.05	100	96.15	**100**
	Avg	95.43	93.28	94.79	76.09	84.90	87.78	**97.45**
50%	CosmosQA	80.29	70.78	**80.70**	32.17	34.72	64.88	76.47
	DREAM	95.10	93.63	93.63	85.20	89.41	88.33	**96.67**
	MCScript	**98.57**	97.85	97.14	94.71	95.99	92.70	**98.78**
	MCScript-2.0	96.40	94.46	96.07	91.02	91.75	91.69	**98.01**
	MCTest	**99.52**	98.81	**99.76**	91.43	95.24	96.19	**99.52**
	QASC	**100**	**99.78**	**99.78**	98.27	98.70	98.49	**100**
	RACE	90.11	87.22	85.23	67.89	71.70	68.18	**93.88**
	RACE-C	85.11	78.09	77.25	66.57	66.85	55.06	**87.36**
	SciQ	**100**	**100**	**99.74**	89.03	96.43	96.43	**100**
	Avg	93.90	91.18	92.14	79.59	82.31	83.55	**94.52**

Table 11: Selective QA accuracies in the multiple choice setting where answer selection is done through ranking before we rank answers for selective QA.

	Dataset	QA+E+C	QA+E	QA+C	E+C	E	C	QA
20%	BioASQ	70.97	70.41	71.55	74.07	74.07	**74.34**	68.99
	HotpotQA	**73.44**	73.08	70.88	71.59	71.51	70.41	69.41
	NaturalQuestions	**85.59**	85.29	85.45	78.46	78.46	80.53	83.27
	SQuAD	96.22	96.45	95.77	83.15	83.09	81.37	**97.15**
	SQuAD-adv	40.39	39.75	39.49	40.07	39.56	**40.59**	31.98
	SQuAD2	35.46	35.24	33.64	36.36	36.13	**36.66**	25.95
	TriviaQA	**64.96**	64.68	64.55	52.67	52.09	52.56	63.98
	Avg	**66.72**	66.41	65.90	62.34	62.13	62.35	62.96
50%	BioASQ	65.96	65.92	64.37	63.53	63.53	**66.95**	64.79
	HotpotQA	64.42	64.21	63.65	65.88	65.85	**66.91**	62.81
	NaturalQuestions	72.28	71.99	70.82	67.54	67.51	**74.18**	69.95
	SQuAD	92.56	**92.57**	92.34	81.86	82.21	80.95	92.54
	SQuAD-adv	33.69	32.90	33.45	**38.74**	38.22	38.52	31.89
	SQuAD2	26.68	25.70	26.00	**32.95**	32.61	32.83	23.52
	TriviaQA	58.40	**58.41**	58.25	51.43	51.18	52.99	58.25
	Avg	**59.14**	58.81	58.41	57.42	57.30	59.05	57.68

Table 12: SelectiveQA on extractive QA using DistillBERT. Note that QA+E+C does best in all cases and contradiction only does second best at 50% coverage.

QA Model	NLI Model	Combination	Confidence	Entailment	Contradiction	Acc
SciQ	mnli-base	QA + C	4.13		-1.06	0.99
		QA + E	3.90	1.37		0.99
		QA + E + C	3.83	1.22	-0.76	0.99
		E + C		2.56	-1.47	0.86
	mnli-large	QA + C	3.98		-1.32	0.99
		QA + E	3.78	1.55		0.99
		QA + E + C	3.65	1.31	-0.97	0.99
		E + C		2.63	-1.72	0.91
RACE	mnli-base	QA + C	3.04		-0.15	0.89
		QA + E	3.03	0.27		0.89
		QA + E + C	3.02	0.26	-0.14	0.89
		E + C		0.73	-0.46	0.75
	mnli-large	QA + C	2.97	0.00	-0.81	0.89
		QA + E	2.91	0.98		0.89
		QA + E + C	2.85	0.92	-0.75	0.89
		E + C		1.76	-1.12	0.78

Table 13: Regression analysis for each mnli-based nli model with each QA model used calibration with logistic regression multiple choice settings. Accuracy is the evaluation metric used.

		Contradiction		Entailment		Neutral	
Dataset	QA	Score	Class	Score	Class	Score	Class
CosmosQA	**0.53**	-0.34	-0.17	0.05	-0.01	0.21	0.16
DREAM	**0.72**	-0.57	-0.35	0.54	0.50	-0.11	-0.13
MCScript	**0.80**	-0.59	-0.42	0.59	0.50	-0.04	-0.08
MCScript2	**0.77**	-0.50	-0.32	0.41	0.37	-0.04	-0.05
MCTest	**0.73**	-0.65	-0.47	0.64	0.69	-0.20	-0.15
QASC	0.57	-0.54	-0.28	0.55	**0.67**	-0.50	-0.26
RACE	**0.65**	-0.37	-0.20	0.35	0.34	-0.11	-0.11
RACE-C	**0.59**	-0.24	-0.13	0.18	0.25	-0.09	-0.11
SciQ	**0.75**	-0.69	-0.47	0.68	0.67	-0.42	-0.19

Table 14: Correlation analysis (Spearman rank correlation) per dataset in the multiple choice setting. RoBERTa-RACE is used for the QA scores.

		Contradiction	Entailment	Neutral	QA
multiple choice	Score	**-0.47**	0.37	-0.06	0.71
	Class	-0.28	**0.38**	-0.06	
extractive QA	Score	-0.16	**0.31**	-0.12	0.19
	Class	-0.15	**0.39**	-0.29	

Table 15: Correlation analysis (Spearman rank correlation) in the multiple choice and extractive QA settings. RoBERTa-RACE is the QA model used for multiple choice QA scores and BERT-large is used for the extractive QA scores.

less correlation than entailment has with the correct answer. Despite the results presented above, contradiction only has a notable correlation with the correct answer when the score is used rather than the classification. This is another proof point of our approach of using confidence scores for NLI rather than classifications. Interestingly in the extractive QA case the neutral class is more negatively correlated when selecting for contradiction when using classification. Our conjecture would be that in the extractive QA case we don't have much to compare against. When looking at the per dataset correlations for the multiple choice setting (Table 14) we see that in most cases, other than the QA confidence scores, the contradiction scores have the strongest correlations with the correct answer out of any NLI class and neutral, as we would expect, tends to have very weak correlations. We do not present the per dataset correlation for extractive QA as they are very weak, which we again hypothesize comes from having no answers to compare with.