Using contradictions improves question answering systems

Anonymous ACL submission

Abstract

This work examines the use of *contradiction* in natural language inference (NLI) for question answering (QA). Typically, NLI systems help answer questions by determining if a potential answer is *entailed* (supported) by some background context. But is it useful to also determine if an answer contradicts the context? We test this in two settings, multiple choice and extractive QA, and find that systems that incorporate contradiction can do slightly better than entailment-only systems on certain datasets. However, the best performances come from using contradiction, entailment, and QA model confidence scores together. This has implications for the deployment of QA systems in domains such as medicine and science where safety is an issue.

1 Introduction

Safety in NLP systems is unresolved, particularly in biomedical and scientific contexts where hallucination, overconfidence, and other problems are major obstacles to deployment (Ji et al., 2022; Kell et al., 2021). One active area of research to solve these issues is natural language inference (NLI) (Li et al., 2022). NLI is the task of determining whether a hypothesis is true (*entailed*), false (*contradicted*), or undetermined (*neutral*) given some premise.

Current NLI systems typically focus only on entailment to verify hypotheses—they calculate the degree to which a hypothesis is supported by the premise. But the premise can provide another signal: contradiction. Regardless of how well a hypothesis is entailed by the context, it can also be more or less contradicted, which could affect whether it is accepted or rejected.

We wondered if adding this signal to a question answering (QA) system might improve performance and safety. To this end, we propose a method that reformulates answers from the QA system as hypotheses for NLI, calculates the entailment, contradiction, and neutrality of each hypothesis, and then selects the best one based on a combination of these results . We show that across 16 QA datasets (9 multiple choice and 7 extractive), the best approach is to use entailment, contradiction, and confidence scores together. Using only contradiction is roughly on par with, and sometimes better than, using only entailment.

1.1 Related work

NLI for question answering has been explored by several authors in various settings; see Paramasivam and Nirmala (2021) for an overview.

One of these settings is **selective question answering for extractive QA**, where *selective* refers to abstention when the system is not confident enough in its answer (Kamath et al., 2020). Chen et al. (2021) have found that NLI systems are able to verify the predictions made by a QA system in this setting, but their result is limited to only selecting a top $k\%$ of answers. Moreover, they do not provide an approach for improving overall performance, nor do they show the effect of incorporating contradiction.

In the related setting of **multiple choice QA and fact checking**, Mishra et al. (2021) have explored the use of entailment, finding that NLI models do well at these tasks by themselves, but can perform even better when they are adapted to in-domain data and longer premises. Yet their method uses only a two-class NLI set up (entailed or not entailed), also ignoring any contradiction signal.

Other QA settings in which NLI has been used include open domain (Harabagiu and Hickl, 2006) and multi-hop (Trivedi et al., 2019). Thus far, all approaches focus only on entailment. To our knowledge, our work is the first to directly leverage contradictions for QA.

Outside of question answering, a domain that uses contradictions is **factual consistency**—the

task of ensuring that a collection of utterances is faithful to a source document. Li et al. (2022) provide an overview. Typically, entailment is still the main focus, but Laban et al. (2022) propose an NLI-based method to ensure the consistency of a summary with a source document using contradiction and neutral scores in addition to entailment, beating out previous systems.

Other researchers have used contradictions to identify consistency errors across Wikipedia (Schuster et al., 2022; Hsu et al., 2021) or generate credible character dialogue (Nie et al., 2021; Song et al., 2020).

2 Methods

We tested the effect of contradictions in two QA settings and a total of sixteen question-answer datasets. Our approach is broadly similar to both Chen et al. (2021) and Mishra et al. (2021).

Briefly, for each dataset, we used pretrained QA models to produce answers and confidence scores for the dataset's questions. We refer to the confidence scores below as **QA**. We then trained QA2D models (where QA2D stands for "question-answer to declarative") to turn the answers into the declarative hypothesis format required for NLI. For example, the question-answer pair "What is the most abundant metal in the Earth crust? Copper." might be rephrased as "The most abundant metal in the Earth crust is copper."

With the question contexts as premises, we then used NLI models to classify every premise-hypothesis pair into three classes, each with an associated score: entailed (**E**), contradicted (**C**), and neutral (**N**). After that, we trained logistic regression calibration models to find which linear combination of the four scores—**QA**, **E**, **C**, and **N**—was best able to pick the answers accurately.

When evaluating performance, we applied the selective QA approach from Kamath et al. (2020) to rank answers using combinations of the four scores, and then consider only those that the model was most confident in answering. We compared selecting the top 20% and 50%. In the multiple choice setting, it was also possible to rank all potential answers according to the four scores, unlike in the extractive QA setting where the QA model produced only one answer per question, so we evaluated performance with that approach as well (see appendix A for details).

3 Experimental setting

In the multiple choice setting, we tested 9 datasets. Two of them are in-domain, since the pretrained QA models we used were finetuned on them. Specifically, we used a RoBERTa large model (Liu et al., 2019) finetuned on the RACE dataset (Lai et al., 2017), as well as two DeBERTa v3 variants, base and xsmall (He et al., 2021a), finetuned on the SciQ dataset (Welbl et al., 2017).

In the extractive QA setting, we used 7 datasets: five from the MRQA 2019 task (Fisch et al., 2019), as well as SQuAD 2.0 (Rajpurkar et al., 2018) and SQuAD adversarial (Jia and Liang, 2017). The SQuAD model is the in-domain dataset: it was used to pretrain (Rajpurkar et al., 2016) the two QA models we used, DistillBERT (Sanh et al., 2020) and BERT-Large (Devlin et al., 2019). Like Chen et al. (2021), we used the Natural Questions dataset for calibration.

In both settings, all datasets contain the relevant context that can be used by the QA models to select answers. More detail on the datasets and QA models is available in appendices B and C respectively.

See appendices D, E, and F for details on the QA2D, NLI, and calibration models. Our models follow the setups described in Kamath et al. (2020), Chen et al. (2021), and Mishra et al. (2021). The main interesting detail is that the calibration models were trained on a holdout set of 100 samples from a single domain, using logistic regression, as in Chen et al. (2021).

4 Results

4.1 Multiple choice setting

For most multiple choice datasets, the best accuracy—when ranking all potential answers—is attained when using a calibrated model combining QA confidence, entailment, and contradiction (**QA+E+C** in Table 1). Only for the in-domain case (RACE-C) does the uncalibrated RoBERTa-RACE model perform on par with that. Using QA scores combined with either entailment (**QA+E**) or contradiction (**QA+C**) achieves similar performance, with contradiction winning by a small margin: 84.33% average accuracy compared to 84.31%.

To inspect these trends further, we performed a correlation analysis of the NLI classes and QA confidence scores with the correct answer (appendix G). We found that besides QA confidence, it is the contradiction score that has the strongest correla-

QA Model	Cosmos	DREAM	MCS	MCS2	MCT	QASC	RACE	RACE-C	SciQ	*Average*
SciQ-base	18.46	43.80	61.99	63.71	44.76	93.41	30.97	27.39	95.28	53.30
SciQ-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 1: *Multiple choice setting*. Accuracy scores (best per column in **bold**, second best underlined) after answer ranking with the mnli-large NLI model. The top three rows show the accuracy of using only the QA models' confidence score; "QA" refers to the scores of the RoBERTa-RACE model, which was used for calibration. The bottom rows add the entailment and/or contradiction scores to the RoBERTa-RACE score. For other NLI models, and for just E, C, and E+C without calibration with RoBERTa-RACE, see Table 8 in the appendix.

tion with the correct answer. The analysis also showed that the neutral class score (**N**) had almost no effect, which is why it is omitted in all results.

When using the selective QA approach and evaluating only the 20% of 50% most confident answers, the best performance is attained with the **QA+C** combination (Table 2). This model is the only one that beats just using the QA confidence score on average. It is stronger than **QA+E+C** and **QA+E** for both coverage percentages.

Contradiction alone, without QA confidence scores (**C**), also beats both entailment alone (**E**) and entailment with contradiction (**E+C**) for both coverages. These results match our intuition that the less contradicted an answer, the more likely it is correct, even in cases where there is uncertainty about its entailment.

4.2 Extractive QA setting

Similar results occur when evaluating the extractive QA datasets with 20% and 50% selective coverage (Table 3). The **QA+C** model does better than **QA** alone, and **C** alone does better than **E+C** or **E** alone, indicating the importance of the contradiction signal here too. However, entailment seems to matter more for extractive QA, as the best F1 score overall was from **QA+E** in the 20% coverage case, and **QA+E+C** in the 50% case.

5 Discussion

Contradiction with background context is a useful signal that NLP systems can use to infer answers to questions. This is not necessarily a superior strategy to using entailment, but our results show that combining these two signals can improve performance beyond what QA models can achieve on their own.

These results are interesting because using contradictions comes with potential benefits for the safety of NLP systems and, as a result, their deployment in domains such as medicine or science. For instance, contradictions help with interpretability: picking the least contradicted answer tells us more about the rejected answers—namely, that they are contradicted by the context—than picking the most entailed one. In addition, once an answer is known to be contradicted, we can try retrieving another answer in an iterative refinement scenario that is difficult to do with entailment. These characteristics suggest lines of inquiry for open domain and generative settings. We know from recent work (Saunders et al., 2022) that self-critici$f$sm is a powerful technique for improving the quality of NLP inference, and future work could assess whether contradiction-based approaches are a valid alternative to the current verification-based ones.

Our work comes with some limitations. It is uncertain whether our results in two specific settings, multiple choice and extractive QA, would extend to more general settings for NLI, although the use of contradictions for factual consistency by Laban et al. (2022) suggests that they could.

Another limitation involves answer ranking and the associated computational cost. The main reason we did not test answer ranking in extractive QA is that we did not generate diverse outputs, but another reason is that such a procedure grows prohibitively expensive as the domain becomes more open. In a fully open domain, ranking would require a quadratic evaluation for each context passage against each reformulated answer candidate (Schuster et al., 2022). Future work should look at comparison approaches that amortize this cost, such as NLI-based dense passage retrieval (Reimers and Gurevych, 2019).

	Dataset	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**
	RACE-C	88.73	92.96	89.44	85.21	85.92	86.62	**93.66**
	SciQ	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	Average	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**
	RACE-C	85.11	84.83	85.39	78.37	78.37	77.25	**87.36**
	SciQ	**100**	**100**	**100**	**100**	**100**	99.74	**100**
	Average	93.90	**94.74**	93.43	91.89	91.24	92.13	94.52

Table 2: *Multiple choice setting*. Accuracy scores (best per row in **bold**, second best underlined) for selective QA with 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
	Natural Questions	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
	Average	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
	Natural Questions	**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
	Average	**68.93**	68.68	68.46	65.39	65.32	66.37	67.76

Table 3: *Extractive QA setting*. F1 scores (best per row in **bold**, second best underlined) for selective QA with 20% and 50% coverage of the dataset. Calibrations and QA confidence are from the BERT-large model, where SQuAD is the in-domain dataset. For similar results on the smaller DistillBERT model, see Table 10 in the appendix.

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.

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.

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].

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.

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.

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.

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].

A Answer ranking procedure

In the multiple choice setting, we performed an answer ranking procedure to pick the answer to a given question among a set of alternative answers N, using both NLI class scores and QA confidence scores. (This is distinct from the selection procedure for the top 20% or 50% of answers we used in both settings.)

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 an NLI class or the QA score, 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:

$$\underset{N}{\operatorname{argmax}}\ \sigma(\mathrm{QA}_n; \mathrm{NLI}_n)$$

where QA_n is the QA model confidence score for answer n, and NLI_n represents the various NLI class scores for n.

We did not use this approach in extractive QA, because we found that asking the model for the top $K = 4$ answer produced almost the same four answer alternatives with slightly different spans each time.

B Datasets

Tables 4 (multiple choice) and 5 (extractive QA) outline the datasets we 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, and MCTest to MCT. The only preprocessing step we performed was to filter out questions where no context passage is provided. Validation splits (as opposed to test splits) are used in the CosmosQA and QASC cases, since context passages or gold standard answers are not available for these datasets.

C QA models

Table 6 outlines the pretrained QA models that we used and the datasets they are trained on. All these models are publicly available on the Hugging Face hub under the locations listed. Where space doesn't allow, RoBERTa-RACE is aliased as RACE.

We trained the two DeBERTa-v3 models (xsmall and base) as shown in Table 7. They were trained using the Hugging Face 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. 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 evaluation suites.

D QA2D models

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. This used the same setup as the DeBERTa-v3 models xsmall and base (see Table 7).

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

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

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).

E NLI models

NLI is used to classify whether the reformulated answer is contradicted, entailed, or neutral with respect to a context passage. We used the whole context, 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). Table 6 contains the Hugging Face references to the NLI models After inference, the confidence scores are used for answer selection and performance evaluation.

E.1 Model size and approach performance analysis

Table 8 mirrors Table 1 in the main text, but shows the accuracy results for uncalibrated **E**, **C**, and **E+C** in the main mnli-large model, as well as the results with the other NLI models, mnli-base and albert-anli. Table 9 shows selective QA accuracy in the multiple choice setting where answer selection is done through ranking before we rank answers for

Hugging Face	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 QA and NLI models 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.

QA Model	Cosmos	DREAM	MCS	MCS2	MCT	QASC	RACE	RACE-C	SciQ	*Average*
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-large										
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
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 8: Accuracy scores in the multiple choice setting for various NLI models used. Calibration was with the RoBERTA-RACE model.

selective QA. Selective QA on extractive QA using DistillBERT (table 10) shows that **QA+E+C** does best in all cases and contradiction only does second best at 50% coverage.

F Calibration models

Like Kamath et al. (2020) and Chen et al. (2021) we developed a set of calibration models in order to perform 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, **C** indicates the contradiction score, and **N** indicates the neutral score. Like in 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.

F.1 Regression Analysis

To illustrate the characteristics of the calibration models, we present a regression analysis for the multiple choice setting (Table 11). 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.

G Correlation Analysis

Since we are using the NLI and QA model scores to construct the setups above, it is useful to know how these factors correlate with the correct answer. Table 13 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 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 a point in favor 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 12), 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.

	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 9: 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
	Natural Questions	**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
	Natural Questions	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 10: 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 11: 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 12: 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 13: 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.