Answer Uncertainty and Unanswerability in Multiple-Choice Machine Reading Comprehension

Anonymous ACL submission

Abstract

Machine reading comprehension (MRC) has drawn a lot of attention as an approach for assessing the ability of systems to understand natural language. Usually systems focus on selecting the correct answer to a question given a contextual paragraph. However, for many applications of multiple-choice MRC systems there are two additional considerations. For multiple-choice exams there is often a negative marking scheme; there is a penalty for an incorrect answer. This means that the system is required to have an idea of the uncertainty in the predicted answer. The second consideration is that many multiple-choice questions have the option of *none of the above* (NOA) indicating that none of the answers is applicable, rather than there always being the correct answer in the list of choices. This paper investigates both of these issues by making use of predictive uncertainty. It is shown that uncertainty does allow questions that the system is not confident about to be detected. Additionally we show that uncertainty outperforms a system explicitly built with an NOA option for the ReClor corpus.

1 Introduction

Machine reading comprehension (MRC), where the correct answer must be deduced for a question from a context paragraph, plays a crucial role in developing systems for natural language processing and understanding. In recent years, popular MRC datasets (Richardson et al., 2013; Chen et al., 2016; Lai et al., 2017; Trischler et al., 2017; Rajpurkar et al., 2018; Yang et al., 2018; Yu et al., 2020) have consistently observed increasingly competitive systems topping public leaderboards (Trischler et al., 2016; Dhingra et al., 2017; Zhang et al., 2021; Yamada et al., 2020; Zaheer et al., 2020) and surpassing human performance. However, systems in deployment should not necessarily always aim to answer a posed reading comprehension question. There are two modes of interest in which an MRC system may choose to abstain from giving an answer: *answer uncertainty* and *unanswerability*. If a system is uncertain about its prediction, it is likely that the predicted answer will be incorrect. In particular, negative marking schemes, which are shown to improve the reliability of multiple-choice assessment as guessing is deterred (Holt, 2006), penalise a system for predicting an incorrect answer while abstaining carries no penalty, and of course the correct answer has a positive reward. In such cases, it would be sensible for a system to abstain from answering if there is answer uncertainty in the prediction. Unanswerability is where the answer to a question is not deducible from the associated context. Consequently, a system should abstain from answering a question if it believes the answer is not present in the context. Answer uncertainty is when the system is unsure about its prediction while unanswerability is where the system (confidently) believes the question cannot be answered.

A fair amount of work has investigated the challenge of tackling unanswerability in span-based reading comprehension (Rajpurkar et al., 2018) with the hope of encouraging systems to truly understand the comprehension task beyond simple word matching with remarkable success (Sun et al., 2018; Hu et al., 2019; Zhang et al., 2021). However, limited work has been completed with regard to unanswerability for multiple-choice reading comprehension datasets, where most work focuses on developing state-of-the-art systems on the default task such as Wan (2020); Jiang et al. (2020). This work investigates both answer uncertainty and unanswerability in multiple-choice MRC.

One challenge for this problem is that unanswerable examples are often not available at training time, and the possible range of incorrect answers even to valid questions is vast. Uncertainty measures have been demonstrated to be effective at out-of-distribution detection across a wide range of tasks (Amodei et al., 2016; Gal, 2016; Malinin,

2019; Malinin et al., 2021). This work studies the potential viability of using uncertainty measures at test time to identify examples for which the system should abstain for both settings of answer uncertainty for optimising performance with a negative marking scheme and handling unanswerability.

2 Multiple-Choice MRC

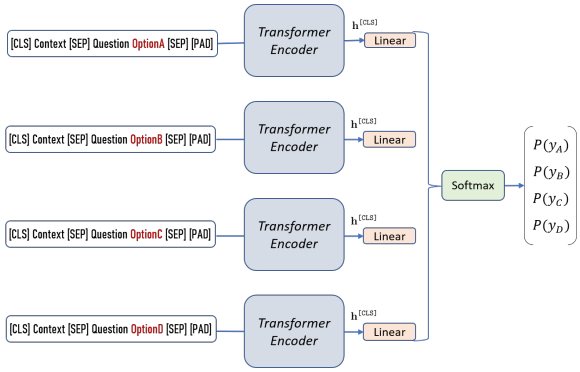


Figure 1: Model architecture.

In the multiple-choice reading comprehension task, the system is given a question, a context passage and multiple possible answer options. The system must be able to select the correct answer option. State-of-the-art for machine comprehension is largely dominated by pre-trained language models (PrLMs) (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019; Lan et al., 2020; Clark et al., 2020; Radford and Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020; Lewis et al., 2020; Raffel et al., 2020) based upon the transformer encoder architecture (Vaswani et al., 2017). Figure 1 depicts the typical model structure of systems for multiple-choice MRC (Yu et al., 2020). In order to use the transformer architecture, the input to the transformer is constructed as follows [1]:

`[CLS] Context [SEP] Question Option [SEP] [PAD] ...`

The transformer models are usually trained with pairs of sentences separated by the `[SEP]` token. The context is used as the first sentence and the question concatenated with an option is used as the second sentence. The construct is repeated for each of the four options. These four pairs of sentences are passed in parallel to the transformer encoder architecture where the weights are shared for each of the inputs. The hidden state embedding associated with the `[CLS]` token is passed to a final linear head (with a non-linear activation) at the end of the transformer encoder that calculates output scores for each answer option which is then converted to a discrete probability distribution over the four answer options using the Softmax activation. Typically, at test time, the predicted answer option is the one with the greatest probability mass.

The work in this paper focuses on ReClor (A Reading Comprehension Dataset Requiring Logical Reasoning) introduced by Yu et al. (2020) that encourages the development of MRC systems beyond a superficial understanding of the context as the dataset was designed to focus on more challenging logical reasoning questions compared to previous multiple-choice datasets including DREAM (Sun et al., 2019), MCTest (Richardson et al., 2013), ARC (Clark et al., 2018) and RACE (Lai et al., 2017). Results are presented on RACE for comparison against ReClor. Additional numbers are provided on COSMOSQA (Huang et al., 2019) in the Appendix A.3.

The architecture of Figure 1 based on the baseline systems introduced by Yu et al. (2020) is used for simplicity as the focus here is on answer uncertainty and unanswerability. The selected model in this paper deviates from the baseline systems as ELECTRA is specifically selected as the PrLM given that it has been proven to achieve state-of-the-art results in other forms of MRC (Zhang et al., 2021) whilst also being smaller than equivalently competitive ALBERT (Lan et al., 2020) systems.

2.1 Answer uncertainty

In the default setting of multiple-choice reading comprehension task, systems are encouraged to always select one of the available answer options for each of the questions. However, there are many multiple-choice tests, such as the UKMT Senior Mathematics Challenge (Pargeter, 2000), that penalise a candidate for selecting the wrong answer, reward the correct answer and give no penalty for not answering the question. Such scoring systems discourage candidates from guessing if they are not confident about the answer. Similarly, multiple-choice MRC systems must also be able to abstain from giving an answer if there is answer uncertainty present in the prediction. Therefore, it is important to develop robust measures of answer uncertainty where the system chooses to only tackle questions that it is able to answer correctly.

Let the total number of questions in a multiple-choice test be denoted $N = N_{\text{correct}} + N_{\text{wrong}} +$

[1] Other permutations of the context, question and answer options were trialled but they give worse performance.

N_{abstain} where N_{correct}, N_{wrong} and N_{abstain} respectively denote the questions that the system answered correctly, answered incorrectly and abstained from answering. For a penalty, p and reward, r, the overall test score, S, becomes,

$$S = rN_{\text{correct}} - pN_{\text{wrong}} \quad (1)$$

where the aim is to maximise the score. Therefore, the ratio p/r dictates the degree of aggression in the negative marking scheme where a larger ratio encourages a system to abstain from answering a greater number of questions to avoid the harsh penalty of selecting the incorrect answer option.

2.2 Unanswerability

Typically, multiple-choice MRC datasets assume that the question for a given example can be answered using one of the answer options. However, several real multiple-choice tests (Odegard and Koen, 2007) exist where none of the answer options address the posed question in relation to the contextual paragraph. An artificial answer option, *none of the above* (NOA), is usually present in such tests for candidates to be able to indicate the unanswerable questions. Unanswerability is further possible in an educational setting for automatic question generation (Kriangchaivech and Wangperawong, 2019) where new questions are automatically generated. Such question generation systems require a verification stage to automatically filter out the questions that are unanswerable in relation to a passage. Therefore, it is important for MRC systems to detect unanswerable questions and only answer the answerable questions.

In this work, two modes of unanswerability are explored. First, the simple set-up is considered where a multiple-choice MRC system is trained with a mixture of answerable and unanswerable examples and then evaluated on in-domain data that has the same proportion of answerable and unanswerable examples. Second, a more challenging mode of operation is considered where only answerable examples are present at training time but a mixture of answerable and unanswerable examples at test time. In this setting, the MRC model must be able to identify unanswerable examples at test time without encountering any such examples for the learning of its parameters. Hence, the test data is distributionally shifted with respect to the training data. In the first mode, the architecture from Figure 1 can be directly used to handle unanswerability as an additional artificial answer option, NOA, can exist for each example with a positive label for this option for all unanswerable examples.

3 Uncertainty

Research in uncertainty estimation is popular in recent years with model averaging (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Ashukha et al., 2020; Ovadia et al., 2019) as the standard approach. In particular, ensemble-based and sampling-based uncertainty estimates have demonstrated effectiveness for both identifying misclassifications and out-of-distribution inputs (Malinin et al., 2021). This work focuses on ensemble-based approaches for multiple-choice MRC as ensembles consistently outperform single models (Ganaie et al., 2021) and offer interpretable uncertainty estimates.

For multi-class classification, various measures of predictive uncertainty can be calculated using the predicted probability distributions over the classes from each of the ensemble members. Measures of knowledge uncertainty include mutual information, expected pair-wise KL divergence, and reverse mutual information; measure of data uncertainty is the average of the entropy of each predicted distribution (expected entropy); while measures of total uncertainty include (negated) confidence and entropy of the average prediction (Gal, 2016; Malinin, 2019). We present results using the expected entropy as the uncertainty measure for abstaining to answer for both a measure of answer uncertainty in a negative marking scheme and a measure of unanswerability when a system does not encounter unanswerable examples at training time [2]. Formally, expected entropy, $\mathbb{E}[\mathcal{H}]$, for a given input is defined as:

$$\mathbb{E}[\mathcal{H}] = -\frac{1}{K}\sum_{k=1}^{K}\sum_{y} P_{\mathcal{M}_k}(y)\log P_{\mathcal{M}_k}(y) \quad (2)$$

where $P_{\mathcal{M}_k}$ denotes the discrete probability distribution using the the k^{th} model member of an ensemble of size K and $y \in \{A, B, C, D\}$.

4 Data and Experimental Set-Up

All experiments are based upon the ReClor and RACE datasets (Yu et al., 2020; Lai et al., 2017) or their variants. This section discusses how the default datasets are modified to perform experiments

[2] Knowledge uncertainty is theoretically better at out-of-distribution detection but empirical results showed the data uncertainty measure was better for unanswerability.

for answer uncertainty and unanswerability as well as performance criteria.

4.1 Training and evaluation data

	Examples	Ans	Unans
TRN-def	4,638	4,638	0
TRN-mixed	18,552	13,914	4,638
TRN-ans	13,914	13,914	0
DEV-def	500	500	0
DEV-mixed	2,000	1,500	500
EVL-def	1,000	1,000	0

Table 1: ReClor: statistics for data splits.

	Examples	Ans	Unans
TRN-def	87,866	87,866	0
TRN-mixed	351,464	263,598	87,866
TRN-ans	263,598	263,598	0
DEV-def	4,887	4,887	0
DEV-mixed	19,548	14,661	4,887
EVL-def	4,934	4,934	0

Table 2: RACE: statistics for data splits.

Table 11 summarises the statistics for ReClor. Yu et al. (2020) split the ReClor datset into a train, validation and test set that are respectively referred to here as the default (def) configurations: TRN-def, DEV-def and EVL-def. In this default configuration, each example consists of a unique question, contextual paragraph and four answer options with no overlap across the total 7,138 examples in the dataset. All questions have a correct answer amongst the four answer options such that all three default splits are 100% answerable.

Two further training splits are introduced in Table 11 beyond the default configurations: TRN-mixed and TRN-ans. TRN-mixed consists of a mixture of answerable and unanswerable examples, with exactly 25% unanswerability. In contrast, TRN-ans consists of only answerable examples that is 3 times TRN-def. Finally, DEV-mixed is the development set equivalent of TRN-mixed that consists of 25% unanswerable examples too.

Table 2 presents the equivalent statistics and modified datasets for RACE with the main distinction that RACE is a significantly larger dataset.

4.2 Data construction

This section describes the method by which the modified data splits, TRN-mixed, TRN-ans and DEV-mixed, are constructed from the default data splits of ReClor/RACE, TRN-def, DEV-def and EVL-def. As the default configuration only consists of answerable examples, the mixed datasets aim to achieve an equivalent dataset that also contain unanswerable examples. TRN-mixed is constructed from TRN-def as follows:

1. For each example, replicate it 4 times.

2. For each of the four versions of an example, replace one of the answer options with NOA. Ensure a different answer option is replaced for each version of the example.

3. Re-order each example such that NOA is the fourth (D) answer option.

Therefore, TRN-mixed is exactly 4 times the size of TRN-def with 75% answerable and 25% unanswerable examples. Similarly, DEV-mixed is constructed from DEV-def by following the above steps. TRN-ans is the answerable subset of TRN-mixed. Hence, TRN-ans can be considered to only have three answer options as the fourth NOA option is never the correct answer for this dataset.

Note that TRN-mixed and DEV-mixed consist of real unanswerable examples rather than synthetic equivalents. Moreover, the modified construction is not performed on the evaluation set because the unanswerability experiments have to be performed on the development sets as the default test set labels are not publicly available. See Appendix B for details of hyperparameter tuning of models.

4.3 Performance criteria

General performance on any development or evaluation set is reported in terms of accuracy. This is consistent with the performance metric used on the ReClor dataset and leaderboard (Yu et al., 2020).

In order to measure the effectiveness of uncertainty measures at measuring answer uncertainty for negative marking schemes, it is desirable for the uncertainty measure to be correlated with the error-rate. Therefore, the standard approach to assess robustness and uncertainty of error-retention curves (Gal, 2016; Lakshminarayanan et al., 2017; Malinin et al., 2021) is used here. An error retention curve plots a model's mean error over a dataset as measured by the classification error rate with respect to the fraction of the dataset for which the model's predictions are used. The classification error for a given example is 0 if the prediction

matches the label and 1 otherwise. The fraction of the model's predictions to be used is dictated by thresholding the uncertainty measure where all examples are ordered from lowest to highest uncertainty. Ideally, the uncertainty measure should be perfectly correlated in terms of rank-ordering with the error-rate. Hence, it is expected that with an increasing retention fraction, the error rate will increase as increasingly uncertain examples will be retained. Therefore, the area under the retention curve (R-AUC) is used as an appropriate metric to assess the effectiveness of the uncertainty measure for a negative marking scheme where a lower value for R-AUC is indicative of better performance.

The ability to identify unanswerable examples in DEV-mixed is reported using the area under the precision-recall curve and the binary F1 where precision and recall are equally important. For performance on DEV-mixed, in decoding we use:

$$\hat{w} = \begin{cases} \underset{w \neq w_s}{\text{argmax}} \{P(w|x)\} & \text{if } P(w_s|x) > \beta \\ w_s & \text{otherwise} \end{cases} \quad (3)$$

where $\hat{w}$ denotes the predicted class; $P(w|x)$ denotes the discrete probability distribution output by the model over the classes conditioned on the input; w_s denotes the class corresponding to unanswerable (i.e. NOA) and β denotes the threshold that the probability mass assigned to the unanswerable class must exceed in order to be deemed unanswerable. The value of β is swept in order to find the overall performance at different operating points.

5 Results and Discussion

This section discusses the main findings of how the ELECTRA system fares against existing systems on the ReClor dataset and the role of uncertainty measures in using answer uncertainty for tackling negative marking schemes or detecting unanswerable examples for ReClor and RACE. Expected entropy is the chosen uncertainty measure. See the Appendix for other uncertainty measures' results.

5.1 Baseline results

Table 8 presents how the ELECTRA system compares against other PrLMs as well as the DAGN (Huang et al., 2021) and FocalReasoner (Ouyang et al., 2021) too on ReClor. Out of the presented systems, the ELECTRA systems achieve the best accuracy on DEV-def and EVL-def. Note that the best single ELECTRA system achieves an accuracy of 64.2% on EVL-def that out-performs the human performance of 63% achieved by graduate students (Yu et al., 2020). Ensembling boosts performance by 2.9% to 67.1%. Performance on the EVL-def is reported for the best member of the ensemble (on the development set) to avoid multiple submissions to the official leaderboard.

	Model	DEV-def	EVL-def
Paper	Chance	25.0	25.0
	Students	-	63.0
	BERT	53.8	49.8
	XLNet	62.0	56.0
	RoBERTa	62.6	55.6
Others	ALBERT	-	62.6
	DAGN	65.2	58.2
	Focal	66.8	58.9
Ours	ELECTRA	$67.8_{\pm 1.1}$	—
	-max	69.4	64.2
	-ensemble	70.2	67.1

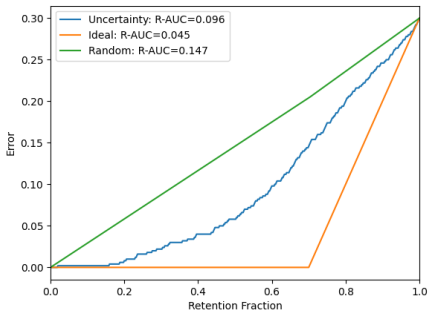
Table 3: Accuracy on default ReClor from the paper Yu et al. (2020); others from the leaderboard and finally our implementations. Mean and standard deviation is quoted for single-seed results.

It is found that pre-training models on RACE (Lai et al., 2017) boosted performance of the best single model to an accuracy of 70.8% on DEV-def and 69.7% on EVL-def. We focus on the situation where only the ReClor data is available for training for fair comparison with other models. At the time of writing, the ELECTRA model ranked 4^{th} on the ReClor leaderboard [3], and only limited details are available for the top three performing systems. However, the focus here is investigating negative marking schemes and unanswerability rather than developing the best system for the ReClor task for which the current system's performance is considered reasonable. See Appendix A.2.1 for the baseline results on RACE. Note, ReClor is considered a significantly more challenging dataset than RACE as human performance on ReClor by graduate students is 63% while human performance on RACE is 94.5%. As the ensembled system achieves superior performance to single systems, the experimental results in the following sections will report results for the ensembled ELECTRA system only.

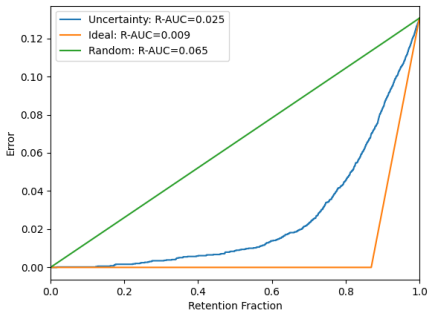
5.2 Answer uncertainty

This section explores the effectiveness of using uncertainty measures for identifying answer uncertainty in the model's predictions to abstain from

[3] Code will be released after anonymity period ends.



(a) ReClor



(b) RACE

Figure 2: Error retention curves for answer uncertainty.

answering for negative marking schemes.

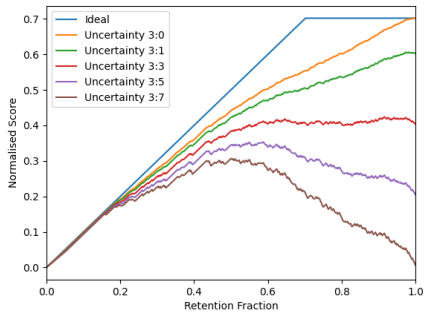
Figure 2 presents the error retention curves for a random measure, an ideal measure and expected entropy as an uncertainty measure for the ELECTRA system trained on TRN-def and evaluated on DEV-def. For ReClor, all curves, as expected, end at a classification error rate of 29.8% when all the data is retained which is consistent with an accuracy of 70.2% from Table 8. The ideal system is where the classification error of each point itself is used as the measure of uncertainty such that all misclassifed points are retained at the end. From Figure 2a, the random system has the largest R-AUC of 0.147 while the ideal system bounds the lowest area at 0.045. The uncertainty measure is able to achieve an R-AUC as low as 0.096 demonstrating that predictive uncertainty measures such as expected entropy are effective at identifying examples that are likely to be misclassified. Similar patterns are observed on RACE from Figure 2b with the main difference that the R-AUC values are lower for all systems as the baseline ELECTRA system on RACE achieves an accuracy of 86.3%. See Appendices A.1.1 and A.2.2 for the R-AUC values for other popular uncertainty measures.

In order to see the impact of using an uncertainty measure for abstaining to answer some questions, Figure 3 illustrates the normalised score using various negative marking schemes while sweeping through the number of examples retained ordered from lowest to highest uncertainty. Each negative marking scheme is expressed as $r : p$ (Equation 1), indicating the reward for a correct answer vs the penalty for an incorrect answer. The normalised score is the total number of points, S, divided by the maximum score achieved by correctly answering all questions. When a harsh negative marking scheme, such as 3:5, is applied it is beneficial to use an uncertainty measure like expected entropy in deployment to filter out the top 40% uncertain examples on ReClor and the top 10% on RACE to achieve the greatest score. Therefore, predictive uncertainty measures help identify examples for which the system should abstain from answering to achieve a higher overall score with aggressive negative marking schemes. However, further work is required to investigate how uncertainty measures may be useful in boosting vanilla performance of answering all questions when using a mild negative marking scheme like 3:1.

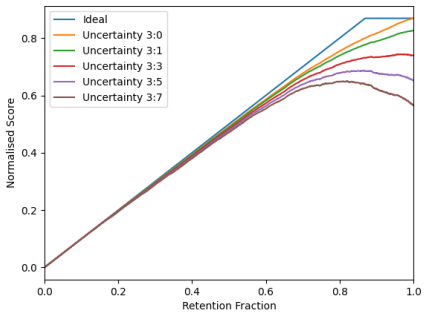
5.3 Unanswerability

Here, we assess the ability of uncertainty measures to identify unanswerable examples in DEV-mixed when using the ensembled ELECTRA-based system. The *Explicit* system trains a four-option system on TRN-mixed (with the fourth option indicative of the question being unanswerable as it corresponds to NOA) while the *Implicit* system trains a three-option system on TRN-ans that contains only answerable examples. This *Implicit* system uses the uncertainty over the three answer options to indicate whether the question is unanswerable. The *Explicit* system takes the maximal probability over the first 3 options and then uses the fourth option probability mass for unanswerability detection by sweeping its value β (Equation 3).

Table 12 presents the best F1 score for each approach at the corresponding precision and recall operating point from the precision-recall curves in Figure 4 for both ReClor and RACE. The area under the precision-recall curve (AUPR) is also reported. As expected, the *Explicit* system is the best performing - with an F1 score and AUPR of 56.0%



(a) ReClor



(b) RACE

Figure 3: Aggressive negative marking schemes.

	Approach	P	R	F1 ↑	AUPR ↑
	Random	25.0	100.0	40.0	25.0
ReClor	Implicit	40.5	63.4	49.5	48.2
	Explicit	50.4	63.0	56.0	55.5
RACE	Implicit	46.1	73.6	56.7	47.9
	Explicit	70.1	70.6	70.3	78.3

Table 4: Detecting unanswerable examples.

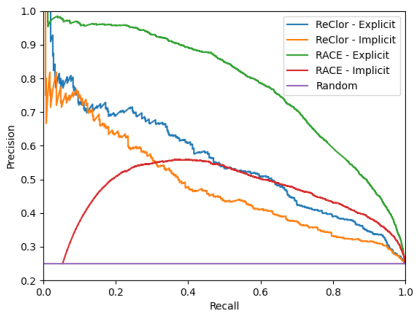


Figure 4: Unanswerabilty detection on DEV-mixed.

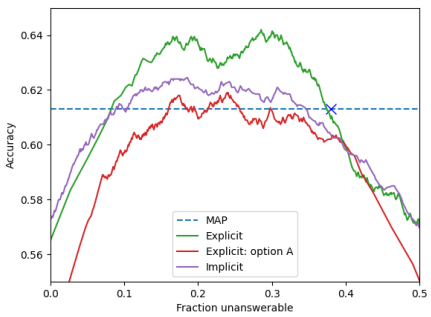
and 55.5% respectively on ReClor, and 70.3% and 78.3% respectively on RACE - as the system encountered unanswerable examples at training time and hence unanswerable examples at test time are in-domain. In contrast, the *Implicit* system did not train with any unanswerable examples. Despite this, the predictive uncertainty, expected entropy in this case, is able to substantially surpass the random system in its ability to detect unanswerable examples at test time to achieve a binary F1 score and AUPR of 49.5% and 48.2% respectively on ReClor, and 56.7% and 47.9% respectively on RACE. Moreover, from the precision-recall curves, the *Implicit* system's ability to identify unanswerable examples surpasses the random curve across all recall rates with the trace lagging behind the *Explicit* system's curve. See Appendix A.1.2 and A.2.3 for the F1 and AUPR scores for other uncertainty measures at detecting unanswerability.

Table 5 compares the *Implicit* and *MAP* system for overall accuracy on DEV-mixed. The maximum-a-posteriori, *MAP*, system is where the ELECTRA system trained on TRN-mixed is directly evaluated on DEV-mixed such that the predicted answer option (out of the four including NOA) is the one with the greatest probability assigned to it. It is interesting to observe that the overall performance of the *Implicit* system at an unanswerability rate of 18.6% is able to outperform the *MAP* system on ReClor. Hence, predictive uncertainty measures are very powerful in this case at identifying unanswerable examples in order to boost overall performance as a system trained on only answerable examples from TRN-ans is capable of out-competing a *MAP* system trained on answerable and unanswerable examples from TRN-mixed. However, the uncertainty measure appears to be weaker on RACE.

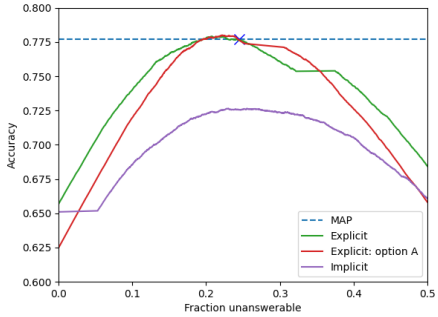
	Approach	ACC ↑	%UNAS
ReClor	Implicit	62.5	18.6
	MAP	61.1	38.0
RACE	Implicit	72.6	23.0
	MAP	77.7	24.5

Table 5: Accuracy (ACC) and Percentage Unanswerable (%UNAS) on Dev-Mixed

Figure 5 shows the performance of the *Implicit* system over a range of thresholds, β, rather than just the maximum performance shown in Table 5. On ReClor, from Figure 5a, it can be seen that it outperforms the MAP decoding over a range of thresholds. However, it is unfair to compare the *Implicit*



(a) ReClor



(b) RACE

Figure 5: Overall performance on DEV-mixed.

system against the *MAP* system alone. Therefore, Figure 5 plots the overall accuracy on DEV-mixed for various systems with a sweep across the number of examples in the dataset predicted as unanswerable. Particularly, the plot for the *Explicit* system is given where the number of examples hypothesised as unanswerable is deduced by sweeping the threshold on the fourth answer option's probability mass (i.e. the probability assigned to NOA) as β. The inference process is as in Equation 3. On ReClor, the *Explicit* system is able to achieve a maximum accuracy of 64.2% at an unanswerability rate of 28.9%. This system outperforms the *MAP* system across a wide range of thresholds of about 10-40%.

As a contrast, the *Explicit: option A*'s performance is also shown. This is generated by sweeping over the threshold on option A rather than the fourth NOA option. If the probability mass assigned to option A is higher than the threshold, the predicted answer will be option A and otherwise the predicted answer is the option with the highest probability mass amongst the other three options. Note, *Explicit: option B* and *Explicit: option C* have similar profiles to *Explicit: option A*. Based on the difference in performance between *Explicit* and *Explicit: option A*, the NOA option operates in a different fashion to the other classes for the ReClor dataset. Intuitively, a possible reason is that the mathematical space for unanswerable questions is a lot larger than the space associated with answerable questions in relation to a specific contextual paragraph which is further evidenced given that the *MAP* system believes 38% of examples are unanswerable despite the unanswerability rate being only 25% at both training and test time.

However, for RACE, from Figure 5b, *MAP* is on par with *Explicit* which in turn peaks with *Explicit: option A*. The inability to out-perform the *MAP* system can be attributed to *MAP* operating at the expected unanswerability rate of about 25%. Therefore, the ability to out-compete a MAP system for ReClor is based on the *MAP* system over-predicting unanswerable examples at decoding time. This tendency to over-predict unanswerable examples may arise due to the complex nature of the questions in ReClor (Appendix C) while other multiple-choice datasets are simpler, leading to a more constrained space learned for NOA.

6 Conclusion

This paper addresses answer uncertainty and unanswerability in multiple-choice MRC. Measures of answer uncertainty are required to identify examples that the system may struggle to get correct and hence should abstain from answering such questions. Unanswerability detection is required for when the answer cannot be deduced using the information provided. An ELECTRA PrLM achieve competitive results on the default ReClor dataset, achieving up to 67.1% accuracy on the evaluation split. Ensemble-based predictive uncertainty measures are explored for both modes of operation: answer uncertainty for negative marking schemes and the presence of unanswerability. It is shown that uncertainty in the prediction such as expected entropy is correlated with the error rate of the MRC system allowing better than vanilla performance with an aggressive negative marking scheme for ReClor and RACE. Interestingly, it is found that expected entropy from the predictions of an implicitly trained system is competitive at unanswerability detection and is able to out-compete MAP decoding from an explicitly trained system that has been trained with unanswerable examples for ReClor.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. http://arxiv.org/abs/1606.06565. ArXiv: 1606.06565.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *ArXiv*, abs/1606.02858.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bhuwan Dhingra, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2017. Linguistic knowledge as memory for recurrent neural networks. *ArXiv*, abs/1703.02620.

Yarin Gal. 2016. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Cambridge.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proc. 33rd International Conference on Machine Learning (ICML-16)*.

M. A. Ganaie, Minghui Hu, M. Tanveer*, and P. N. Suganthan*. 2021. Ensemble deep learning: A review.

Alan Holt. 2006. An analysis of negative marking in multiple-choice assessment. In *19th Annual Conference of the National Advisory Committee on Computing Qualifications (NACCQ 2006)*, pages 115–118.

Minghao Hu, Furu Wei, Yuxing Peng, Zhen Xian Huang, Nan Yang, and Ming Zhou. 2019. Read + verify: Machine reading comprehension with unanswerable questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, abs/1808.05759.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *ArXiv*, abs/1909.00277.

Yin Jou Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. In *NAACL*.

Yufan Jiang, Shuangzhi Wu, Jing Gong, Yahui Cheng, Peng Meng, Weiliang Lin, Zhibo Chen, and Mu li. 2020. Improving machine reading comprehension with single-choice decision and transfer learning.

Kettip Kriangchaivech and Artit Wangperawong. 2019. Question generation by transformers. *ArXiv*, abs/1909.05017.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *EMNLP*.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

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

Andrey Malinin. 2019. *Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment*. Ph.D. thesis, University of Cambridge.

Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark John Francis Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Mariya Shmatova, Panos Tigas, and Boris Yangel. 2021. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *ArXiv*, abs/2107.07455.

Timothy N Odegard and Joshua D Koen. 2007. "none of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory*, 15(8):873–885.

Siru Ouyang, Zhuosheng Zhang, and Hai ying Zhao. 2021. Fact-driven logical reasoning. *ArXiv*, abs/2105.10334.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*.

A Robert Pargeter. 2000. Ukmt yearbook 1998–1999, edited by bill richardson. pp. 121.£ 5. 1999. isbn 0 9536823 0 7 (ukmt, mathematics dept., university of leeds ls2 9jt). *The Mathematical Gazette*, 84(500):344–345.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism.

Fu Sun, Linyang Li, Xipeng Qiu, and Yang P. Liu. 2018. U-net: Machine reading comprehension with unanswerable questions. *ArXiv*, abs/1810.06638.

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.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Rep4NLP@ACL*.

Adam Trischler, Zheng Ye, Xingdi Yuan, Jing He, and Philip Bachman. 2016. A parallel-hierarchical model for machine comprehension on sparse data. *ArXiv*, abs/1603.08884.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Hui Wan. 2020. Multi-task learning with multi-head attention for multi-choice reading comprehension. *CoRR*, abs/2003.04992.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *EMNLP*.

Zhilin Yang, Zihang Dai, Yiming Yang, J. Carbonell, R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *ArXiv*, abs/2002.04326.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *ArXiv*, abs/2007.14062.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *AAAI*.

Pengfei Zhu, Hai Zhao, and Xiaoguang Li. 2020. Duma: Reading comprehension with transposition thinking.

Appendices

A Additional results

The Appendices detail additional results for answer uncertainty and unanswerability detection when using ensembled-based predictive uncertainty. The main paper uses expected entropy as the uncertainty measure of choice. The below sections explore other popular choices of uncertainty measures, including measures of knowledge uncertainty such as mutual information, expected pair-wise KL divergence (EPKL) and reverse mutual information, and also measures of total uncertainty including negative confidence and entropy of expected. The mathematical justifications for each uncertainty measure is motivated by Gal (2016); Malinin (2019).

A.1 ReClor

A.1.1 Answer uncertainty

Uncertainty measure	R-AUC ↓
negative confidence	0.0939
entropy of expected	0.0942
expected entropy	0.0960
mutual information	0.1003
EPKL	0.1018
rev mutual information	0.1028
Ideal	0.0450
Random	0.1470

Table 6: Effectiveness of uncertainty measures for negative marking schemes measured by area under error-retention curves (R-AUC) on ReClor.

A.1.2 Unanswerability

TRN	Measure	F1 ↑	AUPR ↑
	Random	40.0	25.0
mixed	confidence	56.0	55.5
ans	negative confidence	48.3	45.6
	entropy of expected	48.8	47.5
	expected entropy	49.5	48.2
	mutual information	47.4	36.2
	EPKL	47.4	35.0
	rev mutual information	47.4	34.5

Table 7: Effectiveness of uncertainty measures for unanswerability detection for ReClor.

A.2 RACE

This section details additional results on RACE including the baseline results and comparisons with the other popular choices of uncertainty measures.

A.2.1 Baseline

	Model	DEV-def	EVL-def
Others	Roberta	—	83.2
	ALBERT	—	86.5
	-ensemble		89.4
	ALBERT + DUMA	—	88.0
	-ensemble		89.8
	Megatron-BERT	—	89.5
	-ensemble		90.9
Ours	ELECTRA	$86.5_{\pm 0.3}$	—
	-max	87.0	85.9
	-ensemble	86.9	86.3

Table 8: Accuracy on default RACE. Mean and standard deviation is quoted for single-seed results. Other systems include Roberta (Liu et al., 2019), ALBERT (Lan et al., 2020), ALBERT + DUMA (Zhu et al., 2020) and Megatron-BERT (Shoeybi et al., 2020).

A.2.2 Answer uncertainty

Uncertainty measure	R-AUC ↓
negative confidence	0.0238
entropy of expected	0.0244
expected entropy	0.0246
mutual information	0.0287
EPKL	0.0288
rev mutual information	0.0290
Ideal	0.0085
Random	0.0652

Table 9: Effectiveness of uncertainty measures for negative marking schemes measured by area under error-retention curves (R-AUC) on RACE.

A.2.3 Unanswerability

TRN	Measure	F1 ↑	AUPR ↑
	Random	40.0	25.0
mixed	confidence	70.3	78.3
ans	negative confidence	56.1	46.2
	entropy of expected	56.4	46.4
	expected entropy	56.7	47.9
	mutual information	52.3	41.0
	EPKL	52.2	40.6
	rev mutual information	52.0	40.4

Table 10: Effectiveness of uncertainty measures for unanswerability detection for RACE.

A.3 COSMOSQA

COSMOSQA (Huang et al., 2019) is a multiple-choice reading comprehension dataset that has naturally occurring unanswerable examples. Further results are investigated on this dataset for reference.

A.3.1 Data

	Examples	Ans	Unans
TRN-def	25,262	22,199	3,063
TRN-ans	22,199	22,199	0
DEV-def	2,985	2,541	444
DEV-ans	2,541	2,541	0

Table 11: Statistics for data splits for COSMOSQA.

These numbers disagree with those quoted in the paper in terms of number of samples and in terms of the unanswerability rate suggesting that some data has been modified or removed since the release of the original data. The following results are presented using an ensemble of 5 ELECTRA models, which is consistent with RACE. Expected entropy is used here as the main uncertainty measure.

A.3.2 Unanswerability

Approach	P	R	F1 ↑	AUPR ↑
Random	14.9	100	25.9	14.9
Implicit	50.2	47.1	48.6	52.4
Explicit	71.9	58.3	64.4	72.7

Table 12: Detecting unanswerable examples on default COSMOSQA (DEV-def).

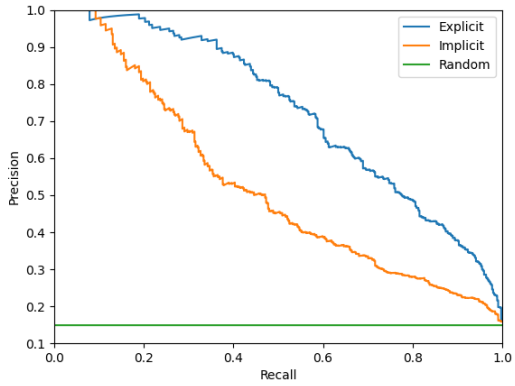


Figure 6: Unanswerabilty detection on DEV-def for COSMOSQA.

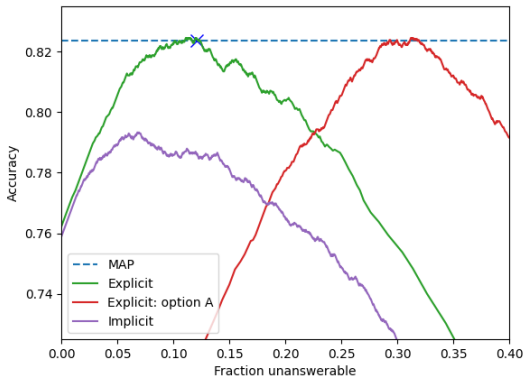


Figure 7: Overall performance on DEV-def for COSMOSQA.

B Hyperparameter tuning

An ensemble of 10/5/5 members for ReClor, RACE and COSMOSQA respectively are trained using the large [4] ELECTRA PrLM as a part of the multiple-choice MRC architecture depicted in Figure 1. Each model has 340M parameters. Grid search was performed for hyperparameter tuning with the initial setting of the hyperparameter values dictated by the baseline systems from Yu et al. (2020). Apart from the default values used for various hyperparamters, the grid search was performed for the maximum number of epochs $\in \{2, 5, 10\}$; learning rate $\in \{2e-7, 2e-6, 2e-5\}$; batch size $\in \{2, 4\}$; truncated length of number of input tokens of the concatenated context, question and a given answer option $\in \{256, 512\}$. For systems trained on ReClor the final hyperparameter settings included training for 10 epochs at a learning rate of 2e-6 with a batch size of 4 and inputs truncated to 256 tokens. For RACE, training was performed for 2 epochs at a learning rate of 2e-6 with a batch size of 4 and inputs truncated to 512 tokens. For COSMOSQA, training was performed for 5 epochs at a learning rate of 2e-6 with a batch size of 4 and inputs truncated to 256 tokens. Cross-entropy loss was used at training time with models built using NVIDIA V100 graphical processing units with training time under 10 hours per model for ReClor, 12 hours for COSMOSQA and 20 hours for RACE. All hyperparameter tuning was performed by training on TRN-def and selecting values that achieved optimal performance on DEV-def. As there is no

[4] Configuration at: https://huggingface.co/google/electra-large-discriminator/blob/main/config.json.

equivalent evaluation set available for the modified versions of ReClor, the final setting of hyperparameters of the system trained on TRN-def is also used for training on TRN-mixed and TRN-ans.

C Examples

This section takes a look at example questions from RACE, COSMOSQA and ReClor to compare the nature of the questions from each dataset.

ReClor

Context:
In a business whose owners and employees all belong to one family, the employees can be paid exceptionally low wages. Hence, general operating expenses are much lower than they would be for other business ventures, making profits higher. So a family business is a family's surest road to financial prosperity.

Question:
The reasoning in the argument is flawed because the argument

Options:

A ignores the fact that in a family business, paying family members low wages may itself reduce the family's prosperity

B presumes, without providing justification, that family members are willing to work for low wages in a family business because they believe that doing so promotes the family's prosperity

C ignores the fact that businesses that achieve high levels of customer satisfaction are often profitable even if they pay high wages

D presumes, without providing justification, that only businesses with low general operating expenses can succeed

Figure 8: Example question from ReClor.

RACE

Context:
This is Jim's room. It's not big, but it's very clean. There is a bed in the room. It's near the door. Under the bed, there are two balls. There is a desk and a chair near the window. There are two pictures in the room, too. They are on the wall.

Question:
Jim's bed is

Options:

A near the door

B near the window

C on the bookcase

D on the wall

Figure 9: Example question from RACE.

COSMOSQA

Context:
Do I need to go for a legal divorce? I wanted to marry a woman but she is not in the same religion, so I am not concern of the marriage inside church. I will do the marriage registered with the girl who I am going to get married. But legally will there be any complication, like if the other woman comes back one day, will the girl who I am going to get married now will be in trouble or is there any complication?

Question:
Why is this person asking about divorce?

Options:

A If he gets married in the church he won't have to get a divorce

B He wants to get married to a different person

C He wants to know if he doesn't like this girl can he divorce her

D None of the above choices

Figure 10: Example question from COSMOSQA.