Impact of Distractors in Multiple Choice Reading Comprehension

Anonymous submission

Abstract

Multiple choice reading comprehension (MCRC) exams are a standard way of assessing candidates for a wide range of language examinations. Despite the ubiquity of these exams, it is difficult to make good questions where: 1) all distractor options remain feasible when only given the question; and 2) there is only one clear solution when both the context passage and question are given. This work examines some standard MCRC datasets with these attributes in mind. We show that even when no contextual passage information is available, deep learning systems can achieve surprisingly high accuracy. This performance is shown to be a result of the choice of distractors, as for some questions the system can use "general knowledge" (from pre-trained models) to answer the question. Furthermore, this work proposes simple metrics that can flag questions where answers can be eliminated without context, as well as highlighting questions where significant information is extracted from the context. These metrics could aid test designers to develop and refine MCRC exam questions that better assess reading comprehension abilities.

1 Introduction

Reading comprehension (RC) exams have become a ubiquitous assessment method to probe how well candidates can read a passage and understand the text's core meaning. They have been used extensively in university entrance exams, for selection of graduate programs, and in a wide range of language competency examinations (Alderson, 2000). A fundamental assumption of RC exams is that to answer any of the questions correctly, one has to read the passage, comprehend its meaning, and identify the relevant information of a given question. This work, however, shows that this assumption may not always hold and that for several standard RC exams, it is possible to perform well without doing any comprehension at all.

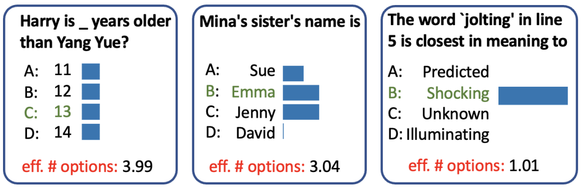

Figure 1: Output probabilities of our model (trained with contexts omitted) on real RACE++ (Liang et al., 2019) examples.

In particular, for several standard multiple-choice machine reading comprehension (MCMRC) datasets (Liang et al., 2019; Huang et al., 2019; Yu et al., 2020), we show that many questions can be answered accurately and confidently without having any access to the contextual passage. Hence, in many cases, it is possible to infer the answer using the question and options alone. Further analysis shows that this is at least partly due to the presence of low-quality distractors, i.e. options that can be eliminated using only the question. As an example, given the question “Mina's sister's name is:", one can eliminate any options that use a traditionally male name (see Figure 1). This highlights a potential ‘shortcut’ that candidates can use for RC exams, where they may be able to answer questions while bypassing the context. Our work raises awareness to this subtle flaw, and demonstrates that systems with degenerate inputs can achieve remarkable performance for standard MCMRC datasets. Further, we propose a simple solution to catch questions that can be answered without comprehension, which could be a useful tool for future multiple-choice RC test designers to ensure that all questions truly assess reading comprehension ability.

Machine reading comprehension (MRC) is a highly researched area, with state-of-the-art (SoTA)

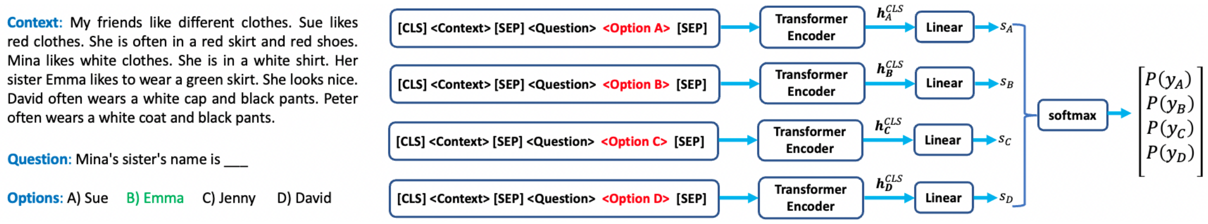
Context: My friends like different clothes. Sue likes red clothes. She is often in a red skirt and red shoes. Mina likes white clothes. She is in a white shirt. Her sister Emma likes to wear a green skirt. She looks nice. David often wears a white cap and black pants. Peter often wears a white coat and black pants.
Question: Mina's sister's name is ___
Options: A) Sue B) Emma C) Jenny D) David
[CLS] <Context> [SEP] <Question> <Option A> [SEP]
[CLS] <Context> [SEP] <Question> <Option B> [SEP]
[CLS] <Context> [SEP] <Question> <Option C> [SEP]
[CLS] <Context> [SEP] <Question> <Option D> [SEP]
Transformer Encoder
h_A^{CLS}
h_B^{CLS}
h_C^{CLS}
h_D^{CLS}
Linear
s_A
s_B
s_C
s_D
softmax
$P(y_A)$
$P(y_B)$
$P(y_C)$
$P(y_D)$

Figure 2: Model architecture.

systems (Trischler et al., 2016; Dhingra et al., 2017; Zhang et al., 2021; Yamada et al., 2020; Zaheer et al., 2020; Wang et al., 2022) often approaching or even exceeding human level performance on public benchmarking leaderboards (Chen et al., 2016; Sun et al., 2019; Richardson et al., 2013; Clark et al., 2018; Lai et al., 2017; Trischler et al., 2017; Yang et al., 2018). Existing work has analysed the robustness of MRC systems: Sugawara et al. (2020) show that for SQuAD (Rajpurkar et al., 2016), systems without questions can be trained to select answer spans better than random. Kaushik and Lipton (2018) demonstrate that systems with degenerate inputs (i.e. using questions or passage only) can achieve reasonable RC performance. Jia and Liang (2017) demonstrate that for RC tasks, adding adversarial sentences to the end of contexts can cause systems to fail, while Si et al. (2019) provide evidence that MCMRC systems don't consider grammar, nor use the whole context, but instead overly focus on particular words. There is also a further larger body of work looking at 'shortcuts' (Geirhos et al., 2020) in natural language processing, where works have analysed system reliance on spurious correlations (Poliak et al., 2018).

Unlike most existing work, the focus of this paper is not in analysing model robustness. Though we do find the surprising result that systems can achieve high MCMRC performance using degenerate inputs, our work uses these shortcut systems for analysis of distractor options which provides insight into the quality of a question. The 4 main contributions of this work are: 1) we are the first investigate whether comprehension is required to answer all questions in MCRC, and highlight this weakness in popular MCMRC datasets. 2) we show that various 'shortcut' systems can achieve surprisingly high performance, and reason that this must be due to poor distractor options. 3) We propose a simple and effective detection technique that can flag low-quality multiple-choice questions - questions that can be answered without considering the context. 4) We provide analysis that looks at the importance of context for MCMRC systems, which further helps in identifying high-quality questions.

2 Multiple choice reading comprehension

Multiple-choice reading comprehension is a common task where given a context passage C and question Q, the correct answer must be deduced from a set of answer options $\{O\}$. Current SoTA MRC systems are dominated by pre-trained language models (PrLMs) based on the transformer encoder architecture (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Lan et al., 2020; Clark et al., 2020). For MCMRC, systems based on ELECTRA (Clark et al., 2020) typically perform amongst the best against comparably sized models (Raina and Gales, 2022), with the standard MCMRC architecture of Figure 2 (Yu et al., 2020; Raina and Gales, 2022). Each option is individually encoded along with the question and the context into a score, and a softmax layer converts the 4 options scores into a probability distribution. At inference, the predicted answer is the option with the greatest probability.

A requirement for good MCRC questions is that information from both the question and the context passage must be used to determine the correct answer. If the answer can be deduced without the context, then this implies that comprehension is not required. To probe whether this is an issue for MCMRC, we propose to train various systems that use degenerate inputs. In particular, three 'shortcut' systems are considered: a system that uses only the options, a system that uses only the question and options, and a system that uses only the options and the context. The standard set-up (of Section 2) is still followed, however each system has altered inputs, as shown in Figure 3.

Consider the output probability distribution of the predicted answer, $\mathrm{P}(y)$. One can measure the entropy of the distribution, $\mathcal{H}(Y)$, which can be converted into the more interpretable *effective number of options*, $\mathcal{N}(Y)$, a value bounded between 1

Options only - {O}	**[CLS] <Option> [SEP]**
Question and Options - Q+{O}	**[CLS] <Question> <Option> [SEP]**
Option and context - {O}+C	**[CLS] <Context> [SEP] <Option> [SEP]**
Baseline - Q+{O}+C	**[CLS] <Context> [SEP] <Question> <Option> [SEP]**

Figure 3: System inputs for shortcut systems.

and the maximum number of options:

$$\mathcal{N}(Y) = 2^{\mathcal{H}(Y)}, \quad \mathcal{H}(Y) = -\sum_{y \in Y} \mathrm{P}(y) \log_2 \mathrm{P}(y) \qquad (1)$$

For well designed questions, one would expect systems with missing information (i.e. the 'shortcut' models) to have no information of what the answer is. This would correspond to a uniform distribution output (the distribution of maximum entropy), with an effective number of options equal to the total number of answer options. However, if the effective number of options is significantly lower than the total number of answer options, then this implies that the model can somewhat infer the answer without comprehension. To probe how much information is gained by the context, one can additionally look at the mutual information of the context. This looks at how much the entropy decreases between the shortcut system and the baseline system which uses the context.

$$\mathcal{I}(Y;C|Q,\{O\}) = \mathcal{H}(Y|Q,\{O\}) - \mathcal{H}(Y|Q,\{O\},C) \quad (2)$$

3 Experiments

3.1 Setup

		M	H	C	All
TRN	# Q	25,241	62,445	12,702	**100,388**
	# C	6,409	18,728	2,437	**27,574**
DEV	# Q	1,436	3,451	712	**5,599**
	# C	368	1,021	136	**1,525**
EVL	# Q	1,436	3,498	708	**5,642**
	# C	362	1,045	135	**1,542**

Table 1: RACE++ data statistics.

RACE++ (ReAding Comprehension dataset from Examinations) (Lai et al., 2017; Liang et al., 2019) is a MCRC dataset of English comprehension questions for Chinese high school students. The questions assess student's understanding and reasoning abilities, with each question having 4 possible options, one of which is correct based on the passage. The questions are divided into three subsets: RACE-M, RACE-H, and RACE-C which are at a middle school, high school and college level respectively. Table 1 outlines the number of questions and contexts in each split.

For the experiments in this work[1], the baseline ELECTRA model[2] is trained on TRN, tuned on DEV and evaluated on EVL. Additional models are then trained and evaluated using the defective inputs shown in Figure 3. The results in the paper focus on RACE++, but all experiments are repeated for ReClor (Yu et al., 2020) and COSMOSQA (Huang et al., 2019) with very similar findings (see Appendix B). Hyperparameter details are given in C.1. Accuracy is used as the default metric when reporting all performances.

3.2 Results

Training data	M	H	C	All
–	25.00	25.00	25.00	25.00
{O}	40.95	42.14	41.53	41.76
Q+{O}	54.81	57.75	60.31	57.32
{O}+C	66.85	69.15	66.24	68.20
Q+{O}+C	88.09	84.42	81.64	85.01

Table 2: Accuracy on EVL.

Here we compare the performance of the baseline MCMRC system against the shortcut systems. Table 2 shows the baseline (Q+{O}+C) achieves a high accuracy of 85.01% on EVL. Note that the system does worse on harder questions, with a 7% drop in accuracy on college level compared to the middle school level questions.

Alarmingly, Table 2 shows that the shortcut models can achieve high accuracy, significantly outperforming a random system despite missing vital information. In particular, a system trained using only the options ({O}) achieves a surprising accuracy of 41.76%, despite not knowing the question nor the context. Additionally, systems without access to the context (Q+{O}) achieve a respectable 57.32%, highlighting that it is possible for a system (and hence potentially candidates) to achieve reasonable performance without performing any comprehension. Systems with no access to questions ({O}+C) achieve an accuracy of 68.20%, which although high, is less worrying since it is reasonable to expect that some options can be eliminated using context information. The following results will focus on the system where the context is not available (Q+{O}), hereon referred to as the shortcut system.

[1] Code will be made available after anonymity period.

[2] ELECTRA-large from https://huggingface.co/docs/transformers/model_doc/electra

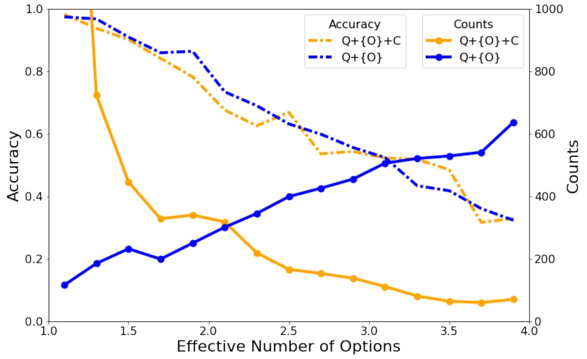

Figure 4: Distribution of effective number of options and corresponding (binned) accuracy.

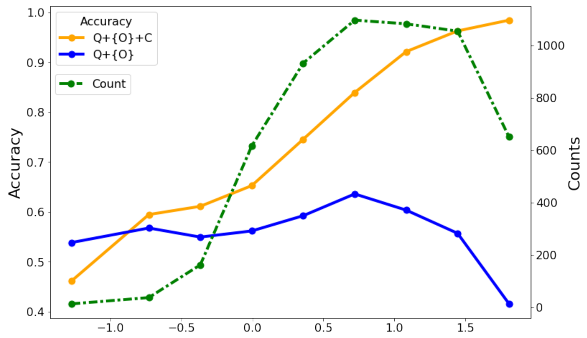

Figure 5: Distribution of counts and corresponding accuracy when points are sorted by MI approximation.

Figure 4 presents the total counts and accuracy (with bin width of 0.2) in terms of the effective number of options (see Equation 1) for both the baseline and shortcut systems. Since the systems are slightly overconfident[3] the systems' output probabilities are calibrated using temperature annealing (Guo et al., 2017) (see Appendix C.2). The solid lines show counts per bin: the baseline system has high certainty for most points, whereas the shortcut model answers some questions confidently and others no more confident than a random guess. There is a strong correlation between accuracy and certainty, where the effective number of options is highly indicative of bin accuracy, showing that the entropy is a very good measure of actual model uncertainty. It is surprising that the shortcut system, without any contextual information, has a significant number of examples in the low entropy region. This shows clearly that for a subset of the questions, the system can confidently reason the answer using only the question and options, due to the presence of disastrous distractors - options that are easy to eliminate using the question alone (see Figure 1). However, this also naturally sheds light on a possible solution, as the shortcut system can be used as a tool to filter out questions which don't require context; questions that have low shortcut entropy[4].

To further look at the influence of context, the mutual information (MI) between prediction and context was approximated for each example using Equation 2. Points with a high MI are questions where the model is certain of the answer with context, but is uncertain without context - a property good questions should have. Figure 5 shows the counts when all the examples are ordered by MI (see Equation 2) along with both the baseline and shortcut system accuracies. In particular, we note that the count distribution has a mix of high and low MI questions, which shows that the benefit of context is not a system-wide property but instead varies over questions. We also note that the accuracy of the baseline system increases considerably when context is useful, while accuracy falls for the shortcut system. It is interesting to note that a small fraction of questions have a negative MI. Though mutual information should always be positive, negative values can be observed since models are only approximations of the true underlying distributions. The low accuracy of the shortcut model on negative MI questions occurs when there is misplaced overconfidence from prior knowledge which is contradicted by information in the context.

4 Conclusions

This work finds that for popular MCMRC datasets, it is possible to achieve high accuracy without performing comprehension. In particular, without vital question information, 'shortcut' systems can still confidently determine the correct answer options and achieve surprisingly high performance. This suggests that answers can be determined without context, and so we propose that these shortcut systems can be used to automatically flag low-quality questions that don't require comprehension. Using an approximation of the mutual information, we also show that the importance of context varies over the questions in the dataset, and reason that high MI questions can be thought of as candidates for high-quality questions that truly measure comprehension abilities.

[3] For both models, the mean of the maximum probability is 5% above the overall accuracy.

[4] The authors carried out an informal evaluation to support this statement. They independently answered without context the 100 lowest and 100 highest entropy examples, achieving an average accuracy of 92% and 32% respectively. All 200 questions and responses are in the supplementary material.

References

J. Charles. Alderson. 2000. *Assessing Reading*, 1 edition. Cambridge University Press,, Cambridge :.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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. 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.

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

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

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.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015.

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

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.

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*, volume 101 of *Proceedings of Machine Learning Research*, pages 742–757, Nagoya, Japan. PMLR.

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. *International conference on machine learning*, abs/1907.11692.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Vatsal Raina and Mark Gales. 2022. Answer uncertainty and unanswerability in multiple-choice machine reading comprehension. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1020–1034, Dublin, Ireland. 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.

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

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does bert learn from multiple-choice reading comprehension datasets? *arXiv preprint arXiv:1910.12391*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

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. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–441.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. 2022. Logic-driven context extension and data augmentation for logical reasoning of text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1619–1629.

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. In *International Conference on Learning Representations*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14506–14514.

Appendix A Limitations

We propose an approach that can automatically flag low-quality questions that can be eliminated without context information. However, it is important to appreciate that questions that are not flagged by such an approach are not necessarily good-quality questions. Further work is required to devise a comprehensive assessment approach for the quality of multiple-choice questions.

Appendix B Additional Results

	TRN	DEV	EVL
RACE++	100,388	5,599	5,642
COSMOSQA	25,262	2,985	–
ReClor	4,638	500	1000

Table Appendix B.1: Overall number of examples per dataset.

Table Appendix B.1 presents the total number of questions in each of the multiple-choice machine reading comprhension datasets RACE++ (Lai et al., 2017), COSMOSQA (Huang et al., 2019) and ReClor (Yu et al., 2020). The baseline systems for each dataset is trained on TRN and hyperparameter tuned on DEV. For RACE++, systems are evaluated on EVL but for COSMOSQA and ReClor, the evaluations are performed on the DEV split because their EVL splits have their labels hidden.

Training data		RACE++	COS.	ReClor
	–	25.00	25.00	25.00
RACE++	{O}	**41.76**	21.44	34.00
	Q+{O}	**57.32**	54.04	34.80
	{O}+C	**68.20**	54.61	46.00
	Q+{O}+C	**85.01**	70.05	48.60
COS.	{O}	29.95	**57.39**	25.20
	Q+{O}	38.73	**68.51**	27.80
	{O}+C	52.41	**78.96**	40.40
	Q+{O}+C	66.81	**84.49**	41.20
ReClor	{O}.	26.07	18.29	**49.00**
	Q+{O}	31.27	33.13	**51.80**
	{O}+C	39.83	36.88	**68.40**
	Q+{O}+C	52.69	41.68	**69.80**

Table Appendix B.2: Cross-performance of systems on RACE++, COSMOSQA (COS.) and ReClor using accuracy.

Table Appendix B.2 shows the cross-performance where systems are trained on one dataset, and evaluated on all 3 (where the same set-up is used). It is evident that in the matched setting, shortcut systems can significantly outperform random guessing across all modes, with similar findings to those in the main paper. It is worth noting that ReClor has a low diversity of questions, and so the options can be quite indicative of the type of question posed, which explains why the questions do not seem to benefit prediction much for ReClor. Additionally, looking at the cross-domain evaluation we see that the systems generalise considerably when applied out-of-domain. Though this is unexpected when viewing the shortcut models as only using spurious correlations, this supports the proposed explanation that the model has existing knowledge that can help identify implausible distractors, independent of the domain.

Appendix B.1 COSMOSQA

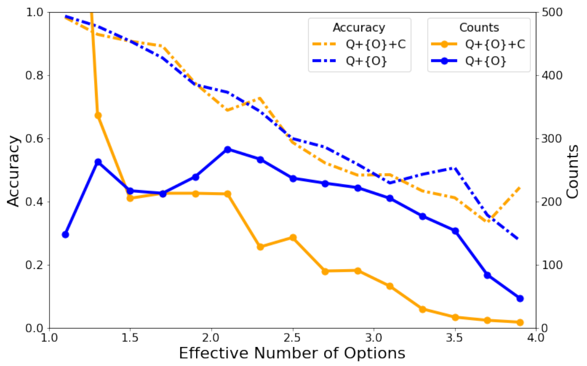

Figure Appendix B.1: Distribution of effective number of options and binned accuracy for COSMOSQA.

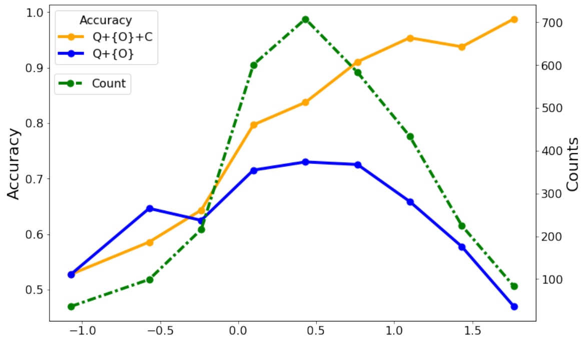

Figure Appendix B.2: Distribution of counts and corresponding accuracy when points are sorted by MI approximation for COSMOSQA.

We repeat the entropy plot (Figure Appendix B.1) for COSMOSQA and find the same similar surprising trend. The entropy of the question is highly indicative of expected accuracy for the point, and further the shortcut system has a very flat distribution showing that a substantial number of examples can be answered confidently and accurately without context. The repeated mutual information plot (Figure Appendix B.2) for COSMOSQA also has the same trend seen in RACE++, validating that our findings are more general that just for RACE++, and holds for the main two standard MCMRC datasets.

Appendix B.2 ReClor

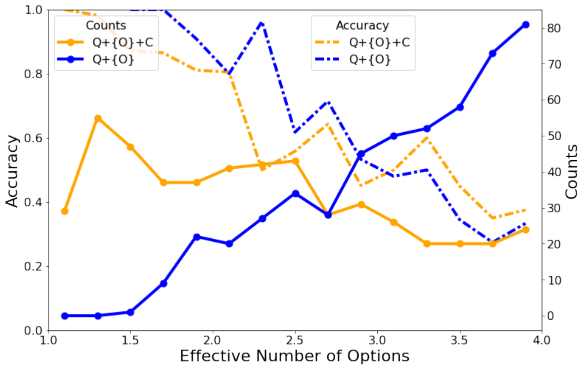

Figure Appendix B.3: Distribution of effective number of options and binned accuracy for ReClor.

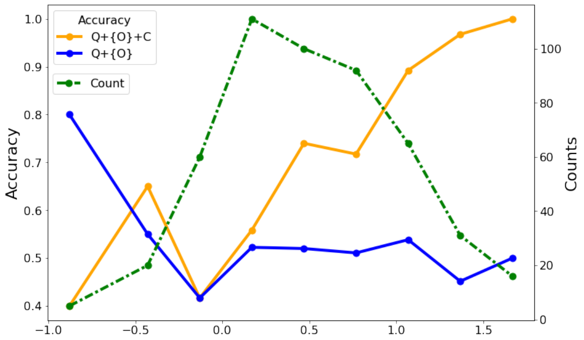

Figure Appendix B.4: Distribution of counts and corresponding accuracy when points are sorted by MI approximation for ReClor.

We generated the same plots for ReClor as well, and observed the same rough general trends. However, the questions of ReClor are much more challenging than in either RACE++ and COSMOSQA, and so we notice that the counts distribution is pushed considerably to the higher entropy side. Additionally, since ReClor is much smaller than RACE++ and COSMOSQA (see Table Appendix B.1), the curves are less smooth and suffer more from noise.

Appendix C Model information

C.1 Training Details

For all systems, deep ensembles of 3 models are trained with the large [5] ELECTRA PrLM as a part of the multiple-choice MRC architecture depicted in Figure 2. 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); Raina and Gales (2022). Apart from the default values used for various hyper-paramters, 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 $\in \{2, 4\}$. 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 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 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 512 tokens. Cross-entropy loss was used at training time with models built using NVIDIA A100 graphical processing units with training time under 3 hours per model for ReClor, 5 hours for COSMOSQA and 4 hours for RACE++. All hyperparameter tuning was performed by training on TRN and selecting values that achieved optimal performance on DEV. For fairness, the 'shortcut' systems (omitting various forms of the input) for each dataset were trained with the same hyperparameter settings as their corresponding baseline systems.

C.2 Calibration

The trained models were calibrated post-hoc using single parameter temperature annealing (Guo et al., 2017). Uncalibrated, model probabilities are determined by applying the softmax to the output logit scores s_i:

$$P(y = k; \boldsymbol{\theta}) \propto \exp(s_k) \tag{3}$$

where k denotes a possible output class for a prediction y. Temperature annealing 'softens' the output probability distribution by dividing all logits by a single parameter T before the softmax.

$$P_{CAL}(y = k; \boldsymbol{\theta}) \propto \exp(s_k/T) \tag{4}$$

As the parameter T does not change the relative rankings of the logits, the model's prediction will be unchanged and so temperature scaling does not affect the model's accuracy. The parameter T is chosen such that the accuracy of the system is equal to the mean of the maximum probability (which would be expected for a calibrated system).

Appendix D Licenses

This section details the license agreements of the scientific artifacts used in this work. The dataset COSMOSQA (Huang et al., 2019) has BSD 3-Clause License. The datasets RACE++ (Lai et al., 2017) and ReClor (Yu et al., 2020) are freely available with the limitation on the latter that it can only be used for non-commercial research purposes. Huggingface transformer models are released under: Apache License 2.0. All the scientific aritfacts are consistent with their intended uses.

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