"World Knowledge" in Multiple Choice Reading Comprehension

Anonymous submission

Abstract

Recently it has been shown that without any access to the contextual passage, multiple choice reading comprehension (MCRC) systems are able to answer questions significantly better than random on average. These systems use their accumulated "world knowledge" to directly answer questions, rather than using information from the passage. This paper examines the possibility of exploiting this observation as a tool for test designers to ensure that the form of "world knowledge" is acceptable for a particular set of questions. We propose information-theory based metrics that enable the level of "world knowledge" exploited by systems to be assessed. Two metrics are described: the expected number of options, which measures whether a passage-free system can identify the answer a question using world knowledge; and the contextual mutual information, which measures the importance of context for a given question. We demonstrate that questions with low expected number of options, and hence answerable by the shortcut system, are often similarly answerable by humans without context. This highlights that the general knowledge 'shortcuts' could be equally used by exam candidates, and that our proposed metrics may be helpful for future test designers to monitor the quality of questions.

1 Introduction

Reading comprehension (RC) exams are used extensively in a wide range of language competency examinations (Alderson, 2000), and have become a ubiquitous assessment method to probe how well candidates can read a passage and understand the text's core meaning. 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. However, recent work has shown that multiple-choice machine reading comprehension (MCMRC) systems without access to the passage can achieve reasonable performance (Pang et al., 2022), showing that the models may be doing less comprehension than anticipated.

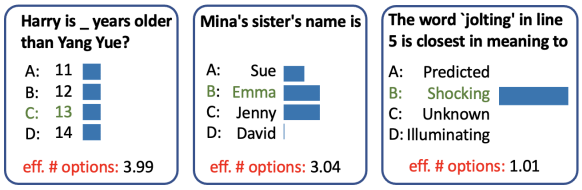

Figure 1: Output probabilities of our model (trained with contexts omitted) on real RACE++ (Liang et al., 2019) examples.

In this paper we analyse this phenomena and for several standard MCMRC datasets (Liang et al., 2019; Huang et al., 2019; Yu et al., 2020) verify that passage-free baselines are able to achieve performance significantly better than random. We show that a subset of questions can be answered accurately and confidently without access to the contextual passage, where further analysis shows this is 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' candidates could use to answer questions while bypassing the context. Our work raises awareness to this potential flaw, and proposes a simple solution to catch questions that can be answered without comprehension. The proposed metrics might 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)

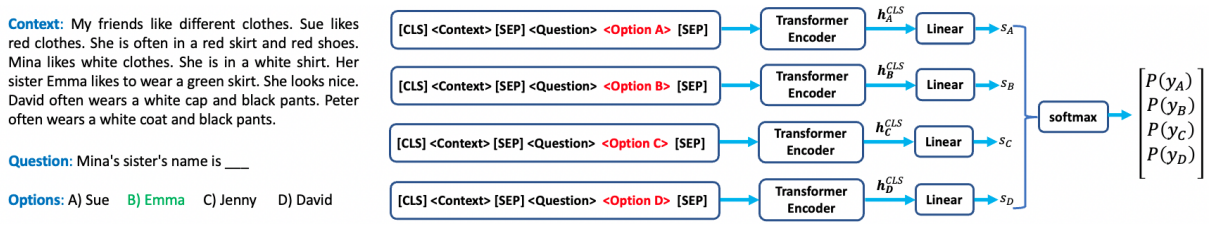

Figure 2: Model architecture.

systems (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 (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, where researchers have questioned whether systems fully leverage context and whether they accomplish the underlying comprehension task (Sugawara et al., 2020; Rajpurkar et al., 2016; Kaushik and Lipton, 2018; Jia and Liang, 2017; Si et al., 2019). Most notably Kaushik and Lipton (2018) show that for a range of question-answering tasks, passage-only systems can often achieve remarkable performance, which has been observed in the MCRC setting (Pang et al., 2022).

Most existing work has discussed model robustness, demonstrating that for some tasks it is possible to obtain high average system performance with no context information. In contrast, this paper focuses on the attributes of individual questions and options, identifying questions where "world knowledge" can be leveraged, and the extent to which this knowledge can be leveraged. This could be a useful tool to enable test designers to monitor the questions being proposed, and whether alternative distractors or questions should be considered.

2 Multiple choice reading comprehension

Multiple-choice reading comprehension is a popular 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; Liu et al., 2019; Clark et al., 2020).

Model Architecture Our question-answering system follows 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.

'No Context' Shortcut System 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. To probe whether this is an issue for MCMRC, we train systems using 'context free' inputs (similar to Pang et al. (2022)). The standard set-up (Figure 2) is still followed, however the input is now altered to exclude the context, as shown in Figure 3.

Standard	Q+{O}+C	[CLS] <Context> [SEP] <Question> <Option> [SEP]
Context-free	Q+{O}	[CLS] <Question> <Option> [SEP]

Figure 3: System inputs for shortcut system.

Effective Number of Options 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 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.

Mutual Information To probe how much information is gained by the context, one can additionally look at an approximation of mutual information of the context. This looks at how much the entropy decreases between the 'no context' shortcut system and the 'context' baseline system .

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

3 Experiments

Data We consider three popular MCMRC datasets: RACE++ (Lai et al., 2017), COSMOSQA (Huang et al., 2019) and ReClor (Yu et al., 2020). RACE++ is a dataset of English comprehension questions for Chinese high school students, COSMOSQA is a large scale commonsense-based reading comprehension dataset, while ReClor is a challenging logical reasoning dataset at a graduate student level. All datasets have 4 options per question, one of which being the correct answer.

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

Table 1: Dataset statistics

Training An ELECTRA-large[1] model is fine-tuned on the training split TRN, hyper-parameters are tuned on the developement set DEV, and performance reported on the test split EVL for RACE++ (DEV splits are used for COSMOS and ReClor due to unavailability of the EVL splits). Additionally, models are trained and evaluated using the 'no context', as described in Section 2. Final hyperparameters are given in Appendix C.1. Three seeds are trained, and ensemble accuracy is used as the default metric when reporting performance.

3.1 Results

Context-Free Performance We compare the performance of the baseline 'standard' system against the shortcut 'no context' systems for the various MCMRC datasets. Table 2 illustrates that the shortcut systems achieve high performance across all MCMRC datasets, all above 50%, significantly above the expected random performance of 25%. Further, we find that the shortcut rules can generalise across domains, most notably seen with the 54% performance when training the shortcut system on RACE and evaluating on COSMOS. This highlights that the shortcut performance cannot be explained purely by dataset bias, but that there is a skill, unrelated to comprehension, that the systems are meaningfully leveraging.

Training data		RACE++	COSMOS	ReClor
	–	25.00	25.00	25.00
RACE++	standard	**85.01**	70.05	48.60
	no context	**57.32**	54.04	34.80
COSMOS	standard	66.81	**84.49**	41.20
	no context	38.73	**68.51**	27.80
ReClor	standard	52.69	41.68	**69.80**
	no context	31.27	33.13	**51.80**

Table 2: Cross-performance of systems on RACE++, COSMOSQA and ReClor using accuracy.

RACE++ Effective Number of Options Figure 4 presents the count and accuracy plots of the effective number of options (bin width of 0.2) for both the standard and shortcut systems on RACE++ (see Appendix for other datasets). Since the systems are slightly overconfident[2], the systems' output probabilities are calibrated using temperature annealing (Guo et al., 2017) (see Appendix C.3).

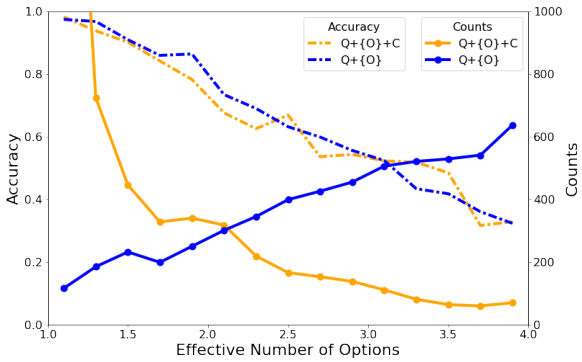

Figure 4: Distribution of effective number of options and corresponding (binned) accuracy.

The baseline system has high certainty for most points, which is somewhat expected given the baseline's high accuracy. However the shortcut system, without any contextual information, has a significant number of examples in the very low entropy region. This shows that for a subset of questions, the system can confidently reason the true answer without doing any comprehension at all. In other cases, the shortcut system can

[1] https://huggingface.co/docs/transformers/model_doc/electra

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

leverage some information from the question and can reduce the number of effective options to between 2-3, which implies that certain poor distractors can be eliminated by the question alone. We also show that for both models, there is a clear linear relationship between uncertainty and accuracy, illustrating that the context-free system's use of world knowledge is sensible and that it leverages meaningful task information. This confirms that the systems are well calibrated and that the effective number of options is a good measure of actual model uncertainty.

Human "World Knowledge" Figure 4 shows that some questions can be answered without any comprehension. To see whether such an insight could be useful as a tool to flag questions that could be improved/replaced, we consider whether such questions are similarly answerable without context by humans. We run an internal human evaluation where we select the 100 questions with lowest and highest entropy and get three graduate students to independently answer the questions in a shortcut setting (without access to the context). We find that humans without comprehension are often able to answer the questions that the shortcut systems is able to confidently answer, with humans achieving an average accuracy of 92% on the 100 lowest and 32% on the 100 highest entropy examples respectively. The sets of questions and graduate student answers are provided in the supplementary information.

	low entropy	high entropy
human	91.7±1.9	31.7±2.9
shortcut	99.0±0.0	24.3±6.2

Table 3: Human and system 'no context' accuracy on lowest and highest entropy questions

Mutual Information 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 questions that require comprehension 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 standard world knowledge is not consistent with the information in the context.

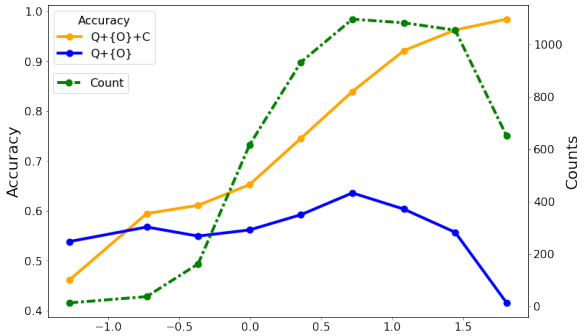

Figure 5: Distribution of counts and corresponding accuracy when points are sorted by MI approximation.

4 Conclusions

For popular MCMRC datasets, systems can achieve reasonably high performance without performing any comprehension. Without passage information, 'shortcut' systems can confidently determine some correct answer options, eliminate some unlikely distractors, and use general knowledge to gain information. Rather than focusing on average system performance, out work analyses individual question's reliance on world knowledge. We propose a metric based on the shortcut systems that may be used to automatically flag questions that are answerable without comprehension. We further provide evidence that the flagged questions are answerable by humans without any context. Lastly, using an approximation of the mutual information, we 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.

References

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

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.

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*.

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.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

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.

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.

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

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, 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 questions that can be answered without contextual information. However, the remaining questions are not necessarily high-quality questions, since many other aspects make up question quality. Further, exams might be aimed at a level where a lack of specific knowledge may be assumed. Our work does not consider variable candidate knowledge levels, and our evaluation was only done by highly educated (we'd like to think) graduate students. Finally, we acknowledge that our human evaluation was limited in size and questions, with potential conflict of self-interest, however it is clearly demonstrated that for low 'shortcut entropy' questions, comprehension is not necessarily required.

Appendix B Additional Results

Appendix B.1 COSMOSQA

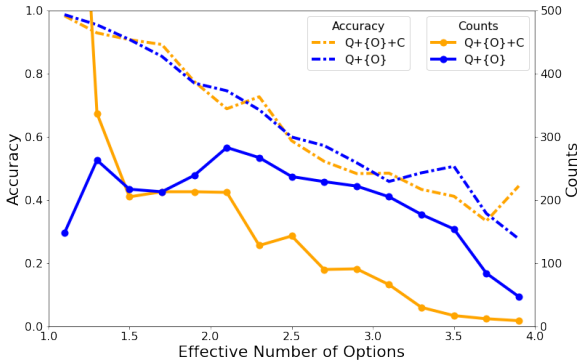

Figure Appendix B.1: Distribution of effective number of options and binned accuracy for COSMOSQA.

We repeat the entropy plot (Figure Appendix B.1) for COSMOSQA and find similar trends to those seen in RACE++. The shortcut no-context system has a very flat distribution with a substantial number of questions answerable without context, with the effective number of options again having

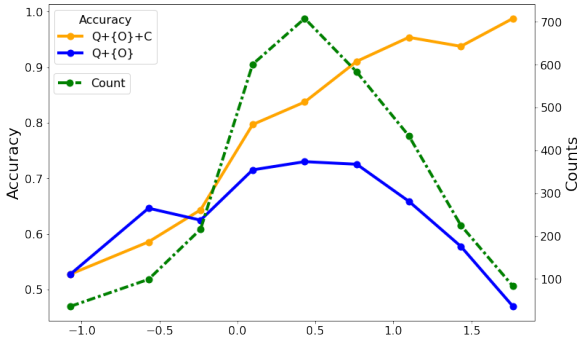

Figure Appendix B.2: Distribution of counts and corresponding accuracy when points are sorted by MI approximation for COSMOSQA.

a clean linear relationship with accuracy. 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++.

Appendix B.2 ReClor

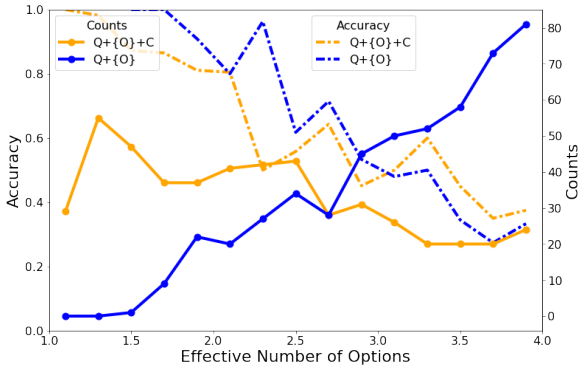

Figure Appendix B.3: Distribution of effective number of options and binned accuracy for ReClor.

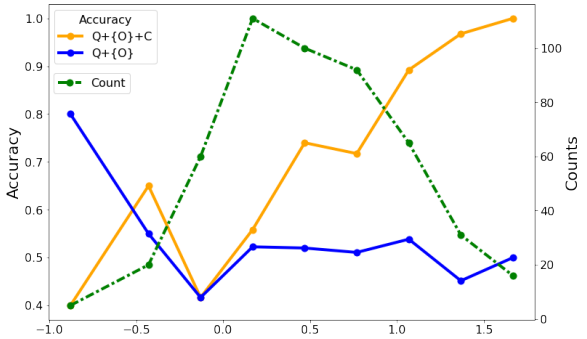

Figure Appendix B.4: Distribution of counts and corresponding accuracy when points are sorted by MI approximation for ReClor.

ReClor show roughly the same trends, however the questions of ReClor are much more challeng-

ing 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 1), the curves are less smooth and largely suffer from noise.

Appendix B.3 Other Shortcuts

We also consider other shortcut approaches, such as having context and options (i.e. missing question) and only options (Figure Appendix B.5). Performance of the systems is shown in Table Appendix B.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 Appendix B.5: System inputs for alternative shortcut systems.

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
COSMOS	{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.1: Cross-performance of systems on RACE++, COSMOSQA and ReClor using accuracy.

Appendix C Model Information

C.1 Training Details

For all systems, deep ensembles of 3 models are trained with the large [3] 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 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 $\in \{2, 4\}$. 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.

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

C.2 Evaluation Details

For each dataset, the systems are trained on the training split and hyperparameter tuned on the development split. For RACE++, systems are evaluated on the held out test data, but for COSMOSQA and ReClor, the evaluations are performed on the development split because their test splits have their labels hidden.

C.3 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.