Distractors for Multiple-choice Questions for Reading Comprehension: Automatic Generation and Evaluation

Yixi Chen*
Department of Information and Library Science
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington
ych21@iu.edu

Shunjie Wang*, **Jiayi Li, John S. Y. Lee**
Department of Linguistics and Translation
City University of Hong Kong
shunjiewang@hotmail.com,
jyli57-c@my.cityu.edu.hk,
jsylee@cityu.edu.hk

Abstract

Reading comprehension is often assessed with multiple-choice questions (MCQs). This paper leverages Large Language Models (LLMs) for automatic generation of distractors for MCQs, and for automatic evaluation of distractor plausibility. We show that fine-tuning an LLM on a modest-sized dataset can lead to automatically generated distractors that rival human-crafted ones in plausibility. Further, we analyze how LLMs differ from humans in making judgments on distractor plausibility.

1 Introduction

Multiple-choice questions (MCQs) are commonly used in assessing reading comprehension (Alhazmi et al., 2024). The quality of an MCQ depends critically on its *distractors*, i.e., the wrong answers that are provided alongside the correct answer. Since it is difficult and time-consuming to craft distractors that seem plausible but are wrong, automating distractor generation (DG) can help alleviate teacher workload and scale up assessment.

This paper leverages Large Language Models (LLMs) for automatic DG and for automatic evaluation of distractor plausibility in MCQs for reading comprehension. While various neural models for DG have been studied (Maurya and Desarkar, 2020; Taslimipoor et al., 2024; Qiu et al., 2020), Large Language Models (LLMs) have so far been evaluated only in zero-shot and few-shot settings on this task. We show that fine-tuning an LLM, even on a dataset of modest size, can produce distractors that can achieve human-level plausibility.

Evaluation is a major challenge in DG research. For traditional automatic metrics such as BLEU, it is difficult to provide references that anticipate all plausible misunderstandings of the passage. As reference-free evaluators, LLMs can offer an efficient alternative, but there has not yet been any reported evaluation on their ability to identify plausible distractors. Specifically, we address the following two research questions:[1]

RQ1: How do distractors generated by a fine-tuned LLM compare to human-crafted ones in terms of plausibility?

RQ2: Can LLMs make accurate judgment on the plausibility of a distractor?

2 Previous Work

Distractor generation (DG). Distractors in MCQs for reading comprehension, unlike those in computer science (Lee et al., 2025) and mathematics (Feng et al., 2024), are derived from possible misunderstandings of the passage provided. Researchers have utilized text-to-text transfer transformers (Taslimipoor et al., 2024), multi-decoder networks (Maurya and Desarkar, 2020), and few-shot approaches with ChatGPT (Bitew et al., 2025). In the only work that attempted fine-tuning LLMs for DG, Offerijns et al. (2020) fine-tuned GPT-2, but did not directly evaluate the plausibility of the generated distractors.

Distractor evaluation. Many DG studies performed automatic evaluation based on lexical and semantic similarity to the gold distractor (Maurya and Desarkar, 2020; Offerijns et al., 2020). However, there are typically not enough references for the variety of distractors that can reflect plausible misunderstandings of the passage. To address this bottleneck in evaluation, the present study assesses the ability of LLMs in making judgments on distractor plausibility.

*Equal contribution.

[1]Data for this paper can be accessed at https://github.com/jikhei/mcq-rc-distractors-inlg2026

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 84–91
October 17-21, 2026 ©2026 Association for Computational Linguistics

Passage		The findings seem to show that children born in the early months of the school year enjoy a double "autumn advantage" – they are already known to have an academic advantage and, now, they also appear to be better equipped for sports, too. The authors believe that autumn-born children's greater exposure over the summer months, towards the end of pregnancy, to vitamin D is the most likely explanation. *John Steele, chief executive of the Youth Sport Trust, said the quality of a young person's introduction to sports at school can be "a major factor" in their sporting development. "Children that get a high-quality first experience, which develops their physical literacy, are those that will have greater agility, balance and coordination, and are more likely to develop an enjoyment of physical activity and excel in sports as they grow up", he said.*
Question		What is one effect of a high-quality first sports experience, according to Steele?
Answer		Enjoyment of physical activity
In-span distractor	Gold	High expectations from the social environment to be good at sports
	Zero-shot	Greater success in other school subjects.
	SFT	Greater participation in team sports.

Table 1: Example MCQ from the dataset, showing the passage (with the critical span italicized), question, answer and the (gold) in-span distractor (Section 3.1); followed by in-span distractors generated by the zero-shot baseline and SFT (Section 3.2)

3 Approach

3.1 Dataset

The OneStopQA dataset (Berzak et al., 2020) contains 486 human-crafted MCQs for 162 short passages. Besides the correct answer, three distractors, reflecting increasing degrees of miscomprehension, are provided for each MCQ:[2]

Category B (In-span) is a distractor based on a misinterpretation of the *critical span*, i.e., the text span in the passage that is relevant to the question.

Category C (Out-of-span) is a distractor derived from a *distractor span*, i.e., a text span that is outside the critical span and is irrelevant to the question.

Category D (Out-of-text) is a distractor based on external knowledge or common sense, without textual support in the passage.

Our study will focus on automatic generation and evaluation of in-span and out-of-span distractors. An example in-span distractor, along with the passage, question and answer from the dataset, is shown in Table 1.

[2]The original scheme used only the letters (B, C, D). The descriptive names are given to facilitate discussion.

3.2 Distractor Generation

Zero-shot (baseline). The LLM is given the passage, question, correct answer, and its *critical span*. It is instructed to:

1. Generate an *in-span distractor* based on the critical span;

2. Select a *distractor span* and then generate an *out-of-span distractor* based on this span.

The prompt is shown in Table 6. Another LLM then verifies whether each generated distractor is an incorrect answer, using the prompt shown in Table 5. If not, the distractor is re-generated.

Supervised Fine-tuning (SFT). The same prompt is used to attempt the same task as the baseline, but the LLM is fine-tuned on the gold distractors from the OneStopQA dataset. Table 1 shows examples of in-span distractors generated by the baseline and by SFT.

3.3 Manual evaluation

Given a passage and a question, human annotators were presented with three distractor candidates — the human-crafted one from the dataset, and two others generated by the zero-shot and SFT methods, respectively. They were then asked to rank them in terms of plausibility based on the following factors:

Textual overlap is the degree of match between the words and phrases in the distractor and in the passage. Similarity in lexical surface forms can make the distractor plausible to the superficial reader (Gao et al., 2019).

Semantic similarity between the distractor and the relevant text in the passage. A plausible distractor can be expected to exhibit only nuanced differences with the original meaning in the passage (Moon et al., 2022).

Logical inference measures the degree to which the distractor is grounded on the passage. A distractor that clearly contradicts the passage would likely appear implausible (Ghanem and Fyshe, 2024; Wang et al., 2023).

3.4 Automatic evaluation

An LLM is instructed to perform the same plausibility ranking task as the humans (Section 3.3), using the prompt shown in Table 7. In an alternative version of the prompt, we presented the three distractor candidates as possible answers and asked the LLM to choose the "best" one. As this version led to lower correlation with the human ranking, the prompt in Table 7 is adopted in the remainder of the paper.

4 Experimental set-up

We randomly sampled 40% (N = 194) of the MCQs in the OneStopQA dataset for the experiment.

Distractor Generation (DG). To facilitate deployment in typical educational settings, we chose a smaller model, `LLama-3.1-8B-Instruct`, to perform DG. We used LoRA (Low-Rank Adaptation) with a rank of 64, trainable projection matrices and 16-bit quantization.[3] Evaluation was based on 9-fold cross-validation.

We chose GPT-4o, a leading LLM outside the Llama family, to check the validity of the generated distractors, i.e., that they were incorrect answers.[4]

Manual Ranking. Two human annotators, both enrolled in a Masters program in Linguistics, participated in the evaluation. Based on the rubric in Section 3.3, they independently ranked the three distractor candidates from #1 (most plausible) to #3 (least plausible), with no ties allowed unless two candidates were identical. Distractors were not ranked if they were judged to be acceptable answers.

Automatic Ranking. GPT-4o was also chosen to perform the ranking task.[5]

Type	Method	% ranked as most plausible		
		Human 1	Human 2	LLM
In-span	Gold	36.08	30.93	30.03
	SFT	**36.60**	**37.63**	32.73
	Zero-shot	31.96	34.54	**37.24**
Out-of-span	Gold	32.47	30.93	24.61
	SFT	**49.48**	**44.85**	35.95
	Zero-shot	18.04	24.22	**39.43**

Table 2: Percentages of gold distractors, distractors produced by the baseline and fine-tuned (SFT) model that were ranked as most plausible by the human annotators; and by the LLM.

5 RQ1: Distractor Plausibility

5.1 Inter-annotator agreement

The annotators achieved a Kendall's τ of 0.4612, which constitutes moderate association (Cohen, 1988), in ranking the plausibility of the in-span distractors (Table 3). In 64.43% of the questions, the same distractor was chosen as the most plausible. While the in-span distractors were based on the same critical span, the out-of-span distractors may be derived from different distractor spans and exhibit more variety in content. For these distractors, the Kendall's τ was 0.3866, which is also considered moderate association (Cohen, 1988). Agreement on the most plausible candidate was 60.82%.

5.2 Results

Table 2 shows how often each of the DG methods produced a distractor that was ranked first in plausibility. The percentages do not necessarily add up to 100% since, for some questions, two methods generated the same distractor.

The "Human" columns report the result of the manual evaluation. For both in-span and out-of-span distractors, the annotators preferred those generated by the SFT model than the zero-shot baseline. This shows that even a small amount of fine-tuning data can yield substantial impact on the DG task.

For out-of-span distractors, the SFT distractors outperformed the gold (human-crafted) ones by 14

[3]Training was performed over 3 epochs with gradient accumulation steps of 8 and mixed precision (bfloat16), at an initial learning rate of 5e-5. All experiments were conducted on a single NVIDIA A100 GPU.

[4]Accessed via Azure OpenAI API, version 2024-05-13.

[5]Accessed via Azure OpenAI API, version 2024-05-13.

Type	Ranking 1	Ranking 2	Kendall's τ
In-span	Human 1	Human 2	**0.4612**
	Human 1	GPT-4o	0.3229
	Human 2	GPT-4o	0.3030
Out-of-span	Human 1	Human 2	**0.3866**
	Human 1	GPT-4o	0.1614
	Human 2	GPT-4o	0.2313

Table 3: Kendall's τ between human annotators and GPT-4o in distractor plausibility ranking

to 17 percentage points.[6] This suggests that LLMs can be on par with humans in choosing an appropriate text span, i.e. the *distractor span*, to craft a convincing answer. The distractors generated by the zero-shot baseline were found to be least plausible.

For in-span distractors, the SFT distractors were as plausible as their human-crafted counterparts for one annotator[7], and slightly more plausible for the other.[8] Unlike the out-of-span distractors, the LLM was provided with the same critical span on which the gold distractor is based (Section 3.2). However, it must conjure a misunderstanding of this span to generate the in-span distractor. The degraded performance thus reflects the challenges in automatic generation of convincing misinterpretations by language learners. This corroborates the difficulties for LLMs in deliberately making flawed inference, as observed in past studies (Peng et al., 2022; Wang et al., 2023).

6 RQ2: Automatic Evaluation on Distractor Plausibility

6.1 Results

As shown in Table 3, GPT-4o yielded weak to modest correlation with human ranking in plausibility, ranging from 0.1614 to 0.3229 in Kendall's τ.

Table 2 compares the differences between GPT-4o and the humans in selecting the most plausible distractor. They agreed that the SFT distractors were more plausible than the gold ones, though the gap was narrower among in-span distractors (32.73% vs. 30.03%) than among out-of-span ones (35.95% vs. 24.61%). However, in contradiction to the humans, GPT-4o found the baseline distractors to be most plausible overall among the three DG methods. This may reflect LLM self-bias, that is, LLM-based systems tend to prefer LLM-presented options (Laurito et al., 2025; Xu et al., 2024; Zheng et al., 2023).

6.2 Correlation analysis

To analyze the discrepancies between human and LLM judgments, we compare the importance they attached to the individual factors in the plausibility rubric (Section 3.3). For each MCQ, we ranked the three distractor candidates in terms of textual overlap, semantic similarity, and logical inference .

Textual overlap was measured by BLEU (Papineni et al., 2002) and ROUGE (uni-gram) (Lin, 2004). The critical span was used as the reference in the case of in-span distractors, and the distractor span in the case of out-of-span distractors.

For both types of distractors, BLEU and ROUGE yielded higher correlation with the human ranking than with the LLM ranking (Table 4). This suggests that the human annotators placed more emphasis on textual overlap than the LLM. They appear to have assessed plausibility from the point of view of weaker readers, who tend to rely on string matching to guess the answer and favor distractors that share similar surface forms as the passage.

Semantic similarity was measured by the token-based metrics BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021)[9], as well as BLEURT (Sellam et al., 2020), which leverages diverse features to mimic human judgment on semantic equivalence.

In terms of BARTScore and BERTScore, the human annotators placed more emphasis on semantic similarity than the LLM. Consistent with the observation above, they seem to have put themselves in the shoes of more elementary readers, who tend to be attracted by distractors that express similar meaning as the passage at a fine-grained, token level. However, the opposite trend was observed for BLEURT, which was more correlated with the LLM ranking. This suggests that the LLM assumed a more advanced reader and emphasized semantic equivalence that may not be expressed by similar surface forms.

Logical Inference was measured with the passage as the premise and the distractor as the hypothesis.[10] A distractor that clearly contradicts the

[6] 49.48% vs. 32.47% for Human 1, and 44.85% vs. 30.93% for Human 2.

[7] 36.60% vs. 36.08% for Human 1

[8] 37.63% vs. 30.93% for Human 2.

[9] Based on the `BART-Large-CNN` model

[10] Using `DeBERTa-Large-NLI` (He et al., 2021) from `https://huggingface.co/microsoft/deberta-large-mnli`

Type	Ranking	Textual overlap		Semantic similarity			Logical inference		
		BLEU	ROUGE	BART-Score	BERT-Score	BLEURT	Entailment	Neutral	Contradiction
In-span	Human 1	**0.159**	**0.233**	**0.237**	0.206	0.157	0.119	0.103	-0.143
	Human 2	0.142	0.223	0.221	**0.212**	0.180	**0.156**	0.103	-0.133
	LLM	0.124	0.150	0.198	0.150	**0.295**	0.023	**0.187**	**-0.243**
Out-of-span	Human 1	0.213	**0.256**	**0.153**	**0.250**	**0.197**	**0.127**	-0.037	-0.002
	Human 2	**0.215**	0.205	0.132	0.247	0.105	0.120	**-0.081**	0.032
	LLM	0.163	0.142	0.092	0.205	0.194	0.097	0.080	**-0.129**

Table 4: Kendall's τ correlation of the metrics in the plausibility rubric (Section 3.3) with human ranking, and with LLM ranking

passage would appear implausible. Hence, the contradiction score is negatively correlated to human and LLM ranking in most settings. For both types of distractors, compared to humans, the LLM more strongly disfavored distractors with high contradiction scores. To improve the correlation with human ranking, the results above suggest that the LLM evaluator can be instructed to place more emphasis on textual overlap than on inference validity.

7 Conclusion

Plausible distractors in MCQs are critical for effective assessment of the test taker's comprehension skills. This paper has leveraged LLMs for automatic DG and automatic evaluation of distractor plausibility. We showed that fine-tuning an LLM on a dataset of modest size can lead to distractors with human-level plausibility, but also revealed that LLMs may be weaker in conjuring misunderstandings of a text. Further, a correlation analysis showed that when making judgments on plausibility, humans tend to place more importance on textual overlap between the passage and the distractor than LLMs.

Our results suggest that future DG research should emphasize the training of LLMs in formulating typical misinterpretations, through in-context learning or fine-tuning. To further improve evaluation accuracy of distractor plausibility, LLMs should be trained or instructed to consider textual overlap as well as the reading level of the intended test taker.

Limitations and Ethical Considerations

The results presented in this paper should be interpreted with the following limitations in mind. First, our evaluation, both manual and automatic, has focused on the plausibility of the distractors. Distractor quality also depends on other factors such as fluency and discriminative power, which should be considered in future research. Second, the training data for fine-tuning the LLM was taken from the same corpus as the test data. To assess whether the improved result can generalize to other domains, the LLM should also be evaluated on passages and questions from other sources.

Acknowledgments

We gratefully acknowledge support from the e-Learning Ancillary Facilities Programme of the Quality Education Fund (Project "Knowledge Overlord - A Self-sustaining AI Game-based Online Platform to Enhance Student's Literacy Ability and 21st Century Skills") and from the University Grants Committee's Teaching Development Grant (Project No. 6000834).

References

Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. 2024. Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14437–14458, Miami, Florida, USA. Association for Computational Linguistics.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

S. K. Bitew, J. Deleu, C. Develder, and T. Demeester. 2025. Distractor Generation for Multiple-Choice Questions with Predictive Prompting and Large Language Models. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases.*

ECML PKDD 2023. Communications in Computer and Information Science, 2134.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Wanyong Feng, Jaewook Lee, Hunter McNichols, Alexander Scarlatos, Digory Smith, Simon Woodhead, Nancy Ornelas, and Andrew Lan. 2024. Exploring Automated Distractor Generation for Math Multiple-choice Questions via Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, page 3067–3082.

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. 2019. Generating Distractors for Reading Comprehension Questions from Real Examinations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6423–6430.

Bilal Ghanem and Alona Fyshe. 2024. DISTO: Textual Distractors for Multiple Choice Reading Comprehension Questions using Negative Sampling. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 6–17.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Walter Laurito, Benjamin Davis, Peli Grietzer, Tomáš Gavenčiak, Ada Böhm, and Jan Kulveit. 2025. AI–AI bias: Large language models favor communications generated by large language models. *Proceedings of the National Academy of Sciences*, 122(31):e2415697122.

Yooseop Lee, Suin Kim, and Yohan Jo. 2025. Generating plausible distractors for multiple-choice questions via student choice prediction. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 23669–23692.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kaushal Kumar Maurya and Maunendra Sankar Desarkar. 2020. Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. In *Proc. 29th ACM international conference on information knowledge management (CIKM)*, page 1115–1124.

Hyeongdon Moon, Yoonseok Yang, Hangyeol Yu, Seunghyun Lee, Myeongho Jeong, Juneyoung Park, Jamin Shin, Minsam Kim, and Seungtaek Choi. 2022. Evaluating the Knowledge Dependency of Questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10512–10526, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. 2020. Better distractions: Transformer-based distractor generation and multiple choice question filtering. In *arXiv:2010.09598*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hsien-Yung Peng, Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2022. Misleading Inference Generation via Proximal Policy Optimization. In *Advances in Knowledge Discovery and Data Mining*, pages 497–509, Cham. Springer International Publishing.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2020. Automatic Distractor Generation for Multiple Choice Questions in Standard Tests. In *Proc. 28th International Conference on Computational Linguistics*, page 2096–2106.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Shiva Taslimipoor, Luca Benedetto, Mariano Felice, and Paula Buttery. 2024. Distractor Generation Using Generative and Discriminative Capabilities of Transformer-based Models. In *Proc. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 5052–5063.

Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang, and Yao-Chung Fan. 2023. Distractor Generation based on Text2Text Language Models with Pseudo Kullback-Leibler Divergence Regulation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12477–12491, Toronto, Canada. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A LLM Prompts

In all prompts, {passage} refers to the passage provided in the MCQ; {question} refers to the question; {answer} refers to the correct answer; and {critical span} refers to the critical span, all taken from the OneStopQA dataset (Berzak et al., 2020).

Text: {passage} Question: {question} Answer: {answer} Based on the text above, is this answer correct (T) or incorrect (F)? Indicate only the letter T or F.

Table 5: Prompt for checking the validity of the distractor, i.e., that it is an incorrect answer (Section 3.2)

The following passage, question, and answer are from a multiple-choice reading comprehension task: Passage: {passage} Question: {question} Answer: {answer} The following is the critical span from the passage on which the correct answer is based: Critical Span: {critical span} The distractor span is another span from the passage that is different from the above critical span. Now, please select such a distractor span from the passage. Please create one distractor based on the given passage, question, answer, and the distractor span you just select. Requirements: 1. The distractor is based on the information in the distractor span, and does not include new information not given in the span. 2. The distractor is similar in length to the given answer. 3. The distractor should be a plausible option to the given multiple-choice question. 4. The distractor must be incorrect; it does not constitute an acceptable correct answer. First, repeat the critical span.
(Ending for in-span distractor generation) Then, respond with the distractor created based on the critical span and nothing else.
(Ending for out-of-span distractor generation) Then, respond with the selected distractor span which is different from the critical span. Lastly, respond with the distractor created based on this distractor span and nothing else.

Table 6: Prompt for generating in-span and out-of-span distractors (Section 3.2)

The following is a reading comprehension passage:
Passage: {passage}

The following is a question from a multiple-choice reading comprehension task:
Question: {question}

The following are three distractors (wrong answers) to the question:
A. {distractor}
B. {distractor}
C. {distractor}

Although the distractors are incorrect, please rank these distractors according to the following criteria:
1. Textual overlap: how many overlaps a distractor shares with the given passage;
2. Semantic similarity: how much similarity a distractor shares with the given passage in terms of semantics;
3. Logical inference: to what degree a distractor is grounded on the given passage;
4. Content concreteness: how many in-text details a distractor includes;

First, evaluate the options one by one according to the above criteria. Give your explanation with no more than 100 words.

Then, according to your rationale, rank the distractors and output the rank using only the letter indices separated by a comma and nothing else.

Table 7: Prompt for ranking distractors in terms of plausibility (Section 3.4). The three distractor candidates, placed in random order, include the gold distractor and the distractors generated by the baseline and SFT models (Section 3.2).