Enhancing Attributed Question Answering using Tailored Progressive Curriculum Learning

Yuhan Chen[1], Bowei Zou[2], Yifan Fan[1], Yuchong Chen[1], Shujun Cao[1], Yu Hong[1]*
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Institute for Infocomm Research, A*STAR, Singapore
{cyhhh1121, yifanfannlp, ycchen0421, tianxianer}@gmail.com
zou_bowei@i2r.a-star.edu.sg; 20245227110@stu.suda.edu.cn

Abstract

We study Attributed Question Answering (abbr., AQA), a newly-released long-form answer generation task. The tailored and efficient training programmes haven't yet been leveraged to strengthen AQA models. This hinders the simultaneous enhancement of their essential capabilities, including evidence identification, cross-source relation recognition and anti-distraction reasoning. To address the issue, we propose a tailored progressive curriculum learning approach, and use it to optimize both encoder-decoder and decoder-only AQA models. Experiments on the benchmark QuoteSum show that our approach yields substantial improvements and enables the AQA performance to reach 73.9% Sem-F1 score. [1]

1 Introduction

For a question, AQA aims to generate a long-form answer that consists of source-dependent evidence and free-style contexts (Bohnet et al., 2022). The evidence is forcibly taken from the given passage-level sources, while free-style contexts can be produced at will for linguistic coherence during generation. Figure 1 shows an example, where evidence is labeled in red font, while contexts the blue.

Recent progress (Gao et al., 2023; Schuster et al., 2024) has been made through end-to-end generation frameworks, which have significantly improved the overall quality of AQA. However, these approaches still face critical limitations as follows:

- The AQA model is required to simultaneously possess the capabilities of 1) evidence identification, 2) cross-source relation recognition, and 3) anti-distraction reasoning.
- Current training strategies typically apply a uniform learning process across these tasks, which lacks dedicated mechanisms to strengthen individual capabilities. Such uniformity hampers the effectiveness of Supervised Fine-Tuning (SFT), particularly when different capabilities demand distinct learning dynamics and granular focus.

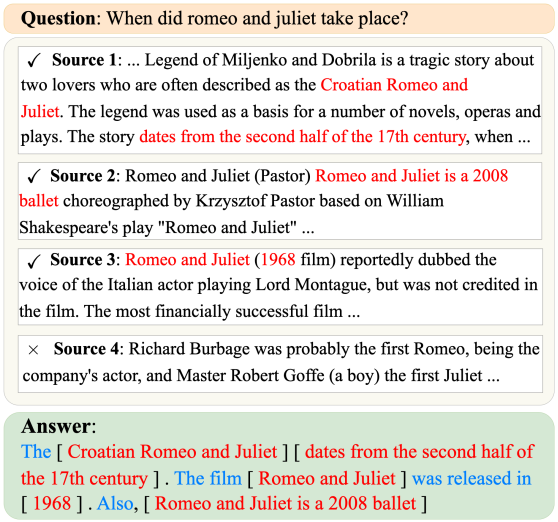

Figure 1: An example of AQA, where the sources labeled with "✓" are reliable cases that comprise certain evidence, while "×" denotes distracting sources.

To address these challenges, we propose a Tailored Progressive Curriculum Learning (TPCL) approach, which introduces a modular and dynamic training strategy. TPCL constructs complexity-controllable training curricula by decomposing AQA into specialized sub-skills and progressively training the model on samples aligned to those sub-skills. Such samples are curated through different tailored training programmes, each focusing on a distinct reasoning or generation aspect. Inspired by curriculum learning theory, TPCL orders the training trajectory from simpler and focused tasks to more complex and integrative ones. This progression not only stabilizes learning but also encourages the model to incrementally acquire and retain AQA capabilities, rather than monolithic SFT.

*Corresponding author
[1]The source code is publicly available at https://github.com/cyh0208/TPCL

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7947–7956
November 4-9, 2025

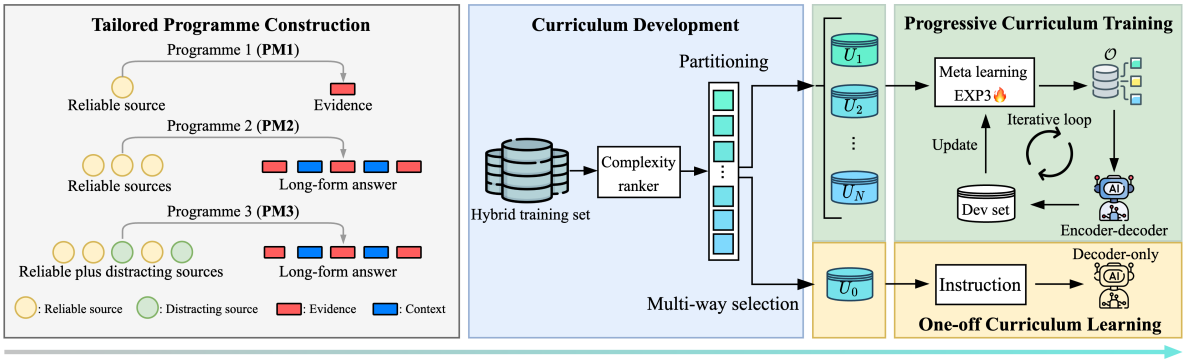

Figure 2: An overview of our proposed method is presented in the figure, where yellow/green circles denote relevant/irrelevant sources, and red/blue squares denote evidence/contexts. The symbol "U" denotes a curriculum.

2 Approach

The TPCL architecture is shown in Figure 2. It comprises three components, including 1) tailored programme construction, 2) complexity based curriculum development, and 3) curriculum learning. Progressive and one-off curriculum learning approaches are respectively proposed to enhance encoder-decoder and decoder-only AQA models.

2.1 Tailored Programme Construction

We construct three tailored training programmes ({PM1, PM2, PM3} for short) towards the enhancement of different capabilities:

- In PM1, an AQA model is fed with the reliable sources S, and it is merely required to predict all the evidence E in S given the question Q. Accordingly, PM1 is used to enhance the capability of evidence identification.
- In PM2, the model is required to analyze the source-wise relationships, and generate the long-form answer $\mathcal{A}$ accordingly. PM2 is used to enhance the capability of grasping the mutually related evidence.
- In PM3, the model is additionally fed with the distracting sources $\bar{S}$ that do not hold any evidence E of $\mathcal{A}$, while it is required to generate $\mathcal{A}$ precisely. PM3 is used to enhance the capability of anti-distraction AQA.

QuoteSum provides explicit labels upon evidence. The labels enable the verification of distracting sources. This allows PM3 to be conducted.

2.2 Curriculum Development

In each n epochs of training (i.e., a learning stage), we intend to provide a curriculum that is not only tailored but controllable in complexity. To meet this requirement, we conduct a curriculum development process as follows:

- We firstly build a hybrid training set which contains the instances derived from different tailored programmes (i.e., PM1-3). Note that each PM1 instance is obtained by filtering evidence out of the long-form answers $\mathcal{A}$, while each PM2 instance is obtained by freezing the distracting sources $\bar{S}$. Each PM3 instance is consistent with the original AQA sample.
- Secondly, we compute the complexity score C_i for each instance in the hybrid training set. We rank all instances accordingly, forming a complexity-aware instance list.
- We divide the above instance list into N partitions, equally and sequentially. The instances in each partition are used as a curriculum for SFT. Thus, we obtain a series of complexity-controllable curricula, each of which, ideally, contains multi-programme instances.

The complexity score C_i is computed in terms of the length l_i of answer, number $\ddot{n}_i$ of sources and distribution density d_i of entities:

$$C_i = \alpha f(l_i) + \beta f(\ddot{n}_i) + \lambda f(d_i) \quad (1)$$

where, f is a normalization function that acts based on the maximum l, $\ddot{n}$ and d occurred in the hybrid training set. The detailed explanations of complexity measurement are presented in Appendix A.

2.3 Conventional Curriculum Learning

Given a curriculum, we divide the instances in it into different batches. Over each batch of data, we conduct SFT by back-propagation. Cross-entropy is used to compute the loss.

On this basis, we use all the N curricula for SFT in the manner of curriculum learning. At each learning stage t ($t \in N$), we follow the common practice (Auer et al., 2002; Matiisen et al., 2019; Bejan et al., 2023) to conduct dynamic curriculum selection, i.e., selecting the most proper curriculum that adapts to the current ability level of the AQA model. It is implemented by EXP3 (Auer et al., 2002) algorithm, where the probability $p_i(t)$ of selecting a proper curriculum is calculated as:

$$\begin{aligned} p_i(t) &= (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{N} w_j(t)} + \frac{\gamma}{N}, \\ w_i(t) &= w_i(t-1) \cdot exp(\gamma \cdot \frac{r_i(t-1)}{p_i(t-1) \cdot N}), \\ r_i(t) &= V(t) + (1 - \frac{Loss(t)}{MaxLoss}) + D(t) \end{aligned} \tag{2}$$

where, $w_i(t)$ serves as the adaptation coefficient imposed upon the i-th curriculum, which is determined by $r_i(t-1)$. $r_i(t-1)$ is a reward given to the model in terms of its performance on the validation set, as well as the complexity of the preceding curriculum. In $r_i(t)$, $V(t)$ denotes the ROUGE-L score, $Loss(t)$ is the cross-entropy loss at stage t, and $D(t)$ is the average difficulty of the training samples in the current curriculum. γ is a smoothing factor. $MaxLoss$ is the cross-entropy loss at the first training stage.

2.4 Progressive Curriculum Learning

We strengthen curriculum learning to avoid 1) unstable transition and 2) catastrophic forgetting.

To ensure a stable transition from easy learning stages to complex ones, we slightly revise EXP3 algorithm. Specifically, we update equation (2) using a dynamic smoothing factor $\gamma(t)$:

$$\gamma(t) = \gamma + \eta\frac{T-t}{T} \tag{3}$$

where, η is an attention factor, while T is the ultimate learning stage (equaling the number of epochs). Accordingly, $\gamma(t)$ pays higher attention to the smoothing factor γ at the earlier learning stage (viz., smaller t). As a result, by equation (2), $\gamma(t)$ helps to randomize the initial curriculum selection process, and thus enables a larger number of easier curricula to be selected for learning.

To alleviate the forgetting of experiences learned from preceding curricula, we conduct a jogging-like curriculum updating. It runs as follows:

- We use EXP3 to select K curricula to initialize a thematic course $\mathcal{O}$ for SFT.
- We update $\mathcal{O}$ at every subsequent learning stage, where k (k<K) newly-selected curricula by EXP3 is used to incrementally replace k existing curricula in $\mathcal{O}$. Progressive learning is always conducted on the updated $\mathcal{O}$.

Backbone	Model	R-L	F1	Rec	Avg
Encoder-decoder framework					
BART-base	+SEMQA	52.7	58.4	77.9	55.5
	+TPCL (Full)	56.4	63.3	78.5	59.7
BART-large	+SEMQA	61.3	69.7	87.1	65.4
	+ICL-SC	60.0	67.9	85.6	63.9
	+TPCL (Full)	62.2	71.5	87.3	66.7
T5-small	+SEMQA	54.3	57.7	68.3	56.0
	+TPCL (Full)	56.9	59.9	77.0	58.4
T5-base	+SEMQA	60.4	67.7	78.0	63.9
	+TPCL (Full)	63.7	71.6	88.1	67.6
T5-large	+ALCE	62.6	71.6	84.4	67.0
	+SEMQA	63.9	71.6	83.7	67.6
	+ICL-SC	59.9	69.2	84.5	64.4
	+TPCL (Full)	**65.7**	**73.9**	**90.0**	**69.2**
Decoder-only framework					
LLaMA-3.1	+SEMQA	42.2	45.1	55.4	43.6
	+TPCL (One-off)	49.0	48.9	56.5	49.0
	+TPCL (Full)	**70.1**	**81.3**	**93.4**	**75.5**
Gemma-3.0	+ALCE	54.1	61.3	79.1	57.6
	+SEMQA	55.9	66.0	79.4	60.7
	+TPCL (One-off)	58.7	67.4	80.5	62.9
GPT-4	+COTAR	68.8	70.5	-	-

Table 1: Performance (%) comparison on Quotesum.

2.5 One-off Curriculum Learning

We simplify TPCL to a one-off version, so as to compatibly couple it with the decoder-only Large Language Models (LLMs). Specifically, given the complexity-aware instance list (mentioned in Section 2.2), we randomly adopt m instances from the list to form the sole curriculum in order. On this basis, we use this curriculum to initiate the demonstration learning within an In-Context Learning (ICL) framework (Dong et al., 2024). Consequently, we prompt LLMs to perform AQA based on the demonstrations. We detail the prompts in Appendix B.

3 Experimentation

3.1 Datasets, Evaluation and Settings

We experiment on QuoteSum (Schuster et al., 2024), which contains 3,131 instances. We divide QutoteSum into training, validation and testing sets in the proportion of 8:1:1, which is the same as the criterion of Schuster et al. (2024). We also follow Schuster et al. (2024) to evaluate all AQA models using the metrics of ROUGE-L (R-L), Sem-F1 (F1) and Sem-Rec (Rec), as well as the geometric mean (Avg) between R-L and F1. We explain the metrics

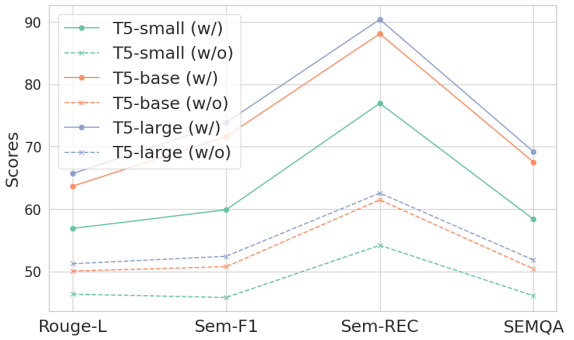

Figure 3: Impact of jogging-like curriculum updating.

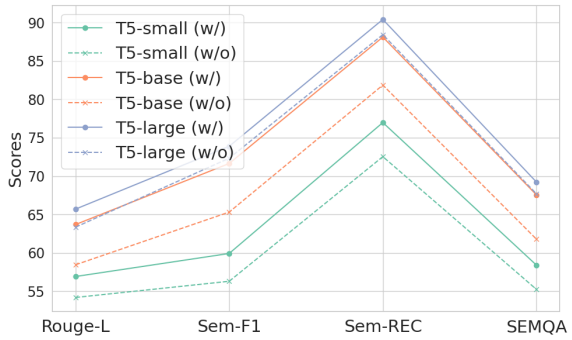

Figure 4: Impact of dynamic smoothing factor $\gamma(t)$.

in Appendix C. The hyperparameters we use comprise those of complexity ranker and curriculum learning, which are presented in Appendix D.

3.2 Main Results and Analysis

Table 1 shows the main results obtained on the test set, where the compared baselines comprise different versions of SEMQA (Schuster et al., 2024) that use a variety of backbones, including BART (Lewis et al., 2020), T5 (Raffel et al., 2020), LLaMA3.1-8B (Grattafiori et al., 2024) and Gemma3-12B (Team et al., 2025). In addition, we compare with CoTAR (Berchansky et al., 2024) and ALCE (Gao et al., 2023), where CoTAR had shown competitive performance for AQA, and ALCE is the essential baseline of SEMQA. We further include ICL-SC (Jia et al., 2022) as the curriculum learning baseline. We overview these models and other related work in Appendix E. Note ALCE and ICL-SC were not evaluated upon QuoteSum for AQA and, therefore, we reproduce them in our experiments.

The test results show that both the full version of TPCL (Section 2.4) and the one-off version (Section 2.5) yield more substantial improvements than the baseline SEMQA, regardless of whether encoder-decoder models or decoder-only LLMs are used as backbones. TPCL consistently outperforms another curriculum learning method, ICL-SC, across all metrics using both BART-large and T5-large backbones. More importantly, the combination of T5-large and TPCL results in a higher F1-score than the strong GPT-4 based CoTAR (73.9% versus 70.5%). This implies that the tailored programmes contribute more to enhancing the abilities of evidence exploration. In addition, the optimized LLaMA-3.1 by TPCL (full) achieves the best performance, though it is more time-consuming than the one-off version.

Backbone	Model	R-L	F1	Rec	Avg
Encoder-decoder framework					
BART-large	+SEMQA	-8.8 ↓	-5.4 ↓	-13.8 ↓	-7.3 ↓
	+TPCL (Full)	-8.5 ↓	-4.8 ↓	-10.9 ↓	-6.9 ↓
T5-large	+SEMQA	-10.8 ↓	-6.5 ↓	-13.9 ↓	-8.8 ↓
	+TPCL (Full)	-10.4 ↓	-8.1 ↓	-13.3 ↓	-8.9 ↓
Decoder-only framework					
LLaMA-3.1	+SEMQA	-3.8	-1.2 ↓	-1.2 ↓	-2.5 ↓
	+TPCL (One-off)	-3.8	+1.7 ↑	+0.9 ↑	-1.2 ↓
Gemma-3.0	+SEMQA	-6.5 ↓	-2.3 ↓	-2.1 ↓	-4.6 ↓
	+TPCL (One-off)	-6.3 ↓	+1.7 ↑	-2.5 ↓	-2.7 ↓

Table 2: Performance degradation (%) on challenge set.

3.3 Ablation Study

We verify the impacts of dynamic smoothing factor $\gamma(t)$ and jogging-like curriculum updating, which are two methods used to consolidate the conventional curriculum learning approach (Section 2.4).

We show the impacts in Figures 3 and 4, where the symbol "w/" denotes the engagement of the two methods, while "w/o" refers to ablation. Different versions of T5 are considered in the ablation study, and TPCL (full) is used as the unabridged framework. The test results show that the two methods have positive impacts all along for all metrics.

3.4 Anti-distraction Verification

To better verify anti-distraction abilities of TPCL and SEMQA, we construct a challenge set. Each instance contains a larger number of paragraph-level sources. GPT-4o (Hurst et al., 2024) is used to produce distracting sources given a deliberately-designed prompt within a self-inspection framework.

We propose a challenging test set by addressing the limitation that most samples in QuoteSum contain only two or three sources (Schuster et al., 2024). To increase the complexity of this test set, we ensure that each sample includes at least one distracting source that is highly relevant but lacks

answer-specific details. Specifically, we construct the challenging test set through the following steps:

- Generation of distracting source: We employ GPT-4o (Hurst et al., 2024) to generate noisy sources that are highly relevant to the topic but lack answer-specific details.

- Verification of generated source: We prompt the model to answer the question using citations from the generated source. If the model can successfully identify the answer within the distracting source, we regenerate the source until no explicit answer can be found. The details of the construction prompts are presented in Appendix F.

- Reordering of sources: Since original sources are typically ordered based on relevance, potentially simplifying the task of answer generation, we randomly reorder the sources provided as input to further increase the difficulty of the test.

Table 2 shows the performance variation compared to that in Table 1, where the mark "-" implies performance degradation occurred on the challenge set, while "+" reflects improvement. Green arrow denotes the less degradation when comparison is made between TPCL and SEMQA, while red the substantial. It can be observed that TPCL yields performance degradation less frequently, exhibiting better anti-distraction capability.

For the challenging test set, performance degradation occurs in almost all models compared to performance on QuoteSum, which indicates that the addition of distracting passages significantly increases the difficulty of question answering. Nevertheless, our method continues to outperform others in most models, demonstrating that our framework enhances the models' anti-interference capability. However, the impact of distractions differs between encoder-decoder models and decoder-only models. Although encoder-decoder models exhibit a clear decline across all metrics, decoder-only models show improvement on certain metrics under similar conditions. This discrepancy can be attributed to the capability of decoder-only LLMs to leverage highly topic-relevant sources, which, despite not directly containing explicit answers, stimulate the retrieval of the models' inherent knowledge to a certain extent.

4 Conclusion

We propose a novel curriculum learning approach (TPCL) for AQA, which provides tailored and complexity-aware curricula for SFT, and avoids unstable transition and catastrophic forgetting during learning. Experiments show that TPCL obtains substantial improvements, and outperforms the previous arts when both encoder-decoder and decoder-only backbones are used. In the future, we will explore a GPT4-oriented ICL method using TPCL.

Limitations

A limitation of our work lies in single-round demonstration ordering within the in-context learning framework. Despite significant performance improvements across various models, TPCL does not specifically explore its potential on larger-sized models (e.g., GPT-4). To solve the limitation, we will develop an iterative curriculum refinement framework where the model autonomously diagnoses curricula through self-improving mechanisms in future research. This enhanced TPCL aims to create a feedback loop between curriculum design and problem-solving states.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under No.62376182.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

Irina Bejan, Artem Sokolov, and Katja Filippova. 2023. Make every example count: On the stability and utility of self-influence for learning from noisy nlp datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10107–10121.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Moshe Berchansky, Daniel Fleischer, Moshe Wasserblat, and Peter Izsak. 2024. Cotar: Chain-of-thought attribution reasoning with multi-level granularity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 236–246.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev,

Jonathan Herzig, and 1 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Tao Fang, Tianyu Zhang, Derek F Wong, Keyan Jin, Lusheng Zhang, Qiang Zhang, Tianjiao Li, Jinlong Hou, and Lidia S Chao. 2025. Llmcl-gec: Advancing grammatical error correction with llm-driven curriculum learning. *Expert Systems with Applications*, page 127397.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Qi Jia, Yizhu Liu, Haifeng Tang, and Kenny Q Zhu. 2022. In-sample curriculum learning by sequence completion for natural language generation. *arXiv preprint arXiv:2211.11297*.

Mingrui Lao, Yanming Guo, Yu Liu, Wei Chen, Nan Pu, and Michael S Lew. 2021. From superficial to deep: Language bias driven curriculum learning for visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3370–3379.

Mike 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. Climb–curriculum learning for infant-inspired model building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 112–127.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. 2019. Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31(9):3732–3740.

Khoi Anh Nguyen, Linh Yen Vu, Thang Dinh Duong, Thuan Nguyen Duong, Huy Thanh Nguyen, and Vinh Quang Dinh. 2025. Enhancing vietnamese vqa through curriculum learning on raw and augmented text representations. *arXiv preprint arXiv:2503.03285*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Tal Schuster, Adam Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William Cohen, and Donald Metzler. 2024. Semqa: Semi-extractive multi-source question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1363–1381.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Nidhi Vakil and Hadi Amiri. 2023. Complexity-guided curriculum learning for text graphs. *arXiv preprint arXiv:2311.13472*.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6095–6104.

Fan Xu, Lei Zeng, Bowei Zou, AiTi Aw, and Huan Rong. 2024. Clffrd: Curriculum learning and fine-grained fusion for multimodal rumor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3314–3324.

Zilin Yuan, Yinghui Li, Yangning Li, Hai-Tao Zheng, Yaobin He, Wenqiang Liu, Dongxiao Huang, and Bei Wu. 2023. A curriculum learning approach for multi-domain text classification using keyword weight ranking. *Electronics*, 12(14):3040.

Question: When was the first women's world cup held?
Answer: [The inaugural tournament was held at a variety of venues across England in June and July 1973] .
Question: Which are the major imports of India?
Answer: Major items of Indian imports [are gold, diamonds, timber, metal scrap] , [crude oil and related products, machinery, electronic goods, gold] , [silver] , [gems, fertiliser, and chemicals.]

Table 3: Two AQA answers of different lengths.

Question: John Simpson is a member of which organization?
Source 1: John Simpson studied architecture at University College London ... He is a member of Royal Institute of British Architects. Simpson is well known for being one of the few modern-day ...
Source 2: He is a founder member of the European Federation of National Institutions for Language and has been a member of its Executive Committee since 2003 ...
Question: When did Romeo and Juliet take place?
Source 1: ... Legend of Miljenko and Dobrila is a tragic story about two lovers who are often described as the Croatian Romeo and Juliet. The legend was used as a basis for a number of novels, operas and plays. The story dates from the second half of the 17th century, when ...
Source 2: Romeo and Juliet (Pastor) Romeo and Juliet is a 2008 ballet choreographed by Krzysztof Pastor based on William Shakespeare's play "Romeo and Juliet" ...
Source 3: Romeo and Juliet (1968 film) reportedly dubbed the voice of the Italian actor playing Lord Montague, but was not credited in the film. The most financially successful film ...
Source 4: Richard Burbage was probably the first Romeo, being the company's actor, and Master Robert Goffe (a boy) the first Juliet ...

Table 4: Two AQA examples with different numbers of sources. Evidence is labeled in red font.

Task instruction:
Answer the QUESTION using the information in the SOURCES. Rules for writing the answer:
1. Copy for every fact.
a. For each factual statement, copy the exact words from the passage. b. Place the copied span in square brackets, immediately followed by its source number, e.g. [exact words][2]. c. If a fact appears in several passages, you may cite multiple source numbers: [exact words][2][5].
2. Maximise explicit copying
a. Use copied spans for all facts whenever possible. b. Add your own words only to link copied pieces together into a fluent sentence.
3. Use all relevant information
a. Include every passage that helps answer the question. b. Do not introduce information that is absent from the passages.
4. Output format
a. Write a single coherent paragraph (or list, if the question requires multiple items). b. Do not repeat the question or list the passages. c. Provide nothing except the final answer.
Shots:
Question: Who sings the song i'm just a love machine?
Source 1: The success prompted her album "Ride to the Rainbow" to be reissued as "Love Machine" for the Japanese release. ...
Source 2: "Love Machine" is a 1975 single recorded by Motown group The Miracles, taken from their album "City of Angels". The song was a 1 Pop smash on the "Billboard" Hot 100, and the biggest-selling hit single of The Miracles' career. ...
Source 3: Girls Aloud performed "Love Machine" on all of their tours and on several live appearances, including at Disney Channel Kids Awards, TMF Awards 2005, and at "The Girls Aloud Party" TV special in 2008. English indie rock band Arctic Monkeys covered the song ...
Answer: [2 "Love Machine" is a 1975 single recorded by Motown group The Miracles, taken from their album "City of Angels".] [3 "Love Machine"] is also the title of a song [3 performed] by the band [3 Girls Aloud] .
... (Other examples)

Table 5: Prompt used for In-context learning with shots.

A Complexity Measurement

To enable curriculum construction, we define a composite complexity score for each training sample, which is computed as the weighted average of three metrics.

Length of Answer. Longer answers typically require more elaborate reasoning. As shown in Table 3, the long answer needs to identify and analyze the relation between evidence spans, while the short answer only identifies.

Number of Sources. More passages increase the complexity of integrating relevant evidence. As shown in Table 4, for the question *When did Romeo and Juliet take place*, the answer synthesizes information from multiple sources: identifying the temporal setting from historical references (e.g., *second half of the 17th century*) and incorporating related content from media adaptations (e.g., *the 1968 film and the 2008 ballet*). It illustrates the dual challenge of attributing facts accurately and distinguishing the distracting source.

Distribution density d_i of entities. We introduce a novel metric to quantify semantic noise based on entity density:

$$d_i = \frac{N_e}{W_P}, \tag{4}$$

where N_e denotes the number of named entities and W_P is the total word count across passages. For entity recognition, we train a RoBERTa-large model (Liu et al., 2019) on ACE05 (Doddington et al., 2004), achieving a 0.94 F1 score.

Parameter	Value	Definition
Complexity score		
α	0.3	weight of answer length
β	0.3	weight of source number
λ	0.3	weight of entities density
Dynamic smoothing factor $\gamma(t)$		
γ	0.1	smoothing factor
η	0.4	attention factor
Encoder-decoder models		
n	3	learning stage
N	6	curriculum number
T	30	ultimate learning stage
K	3	selected curriculums
lr	5e-5	learning rate
Decoder-only models		
m	5	demonstration number

Table 6: The parameters used in our experiment.

B Prompt for In-context Learning

We use prompts to guide LLMs in performing AQA based on demonstrations. Table 5 shows the detailed prompts and shots.

C Metrics

We evaluate all the models with Sem-F1, Sem-Rec and SEMQA metrics proposed by Schuster et al. (2024), which respectively reflect the preciseness of evidence, evidence coverage and overall answer quality. Specifically, Sem-F1 is computed by calculating the normalized token-level F1 between the generated answer and human-written reference answers, with the highest obtained score as the final Sem-F1. Similarly, Sem-Rec is determined by computing the token-level recall between the evidence provided in human-written references and the generated answer, selecting the maximum as the final score. In addition, Rouge-L (Lin, 2004) score is used to compare generated long-form answers with reference answers (Stelmakh et al., 2022; Schuster et al., 2024).

D Hyperparameters

When fine-tuning the encoder-decoder models, we adopt the following hyperparameters. The maximum input length is set to 512 and the batch size to 32. The learning rate is set to 5e-5. For final testing, we select the checkpoint with the highest ROUGE-L on the validation set. Within in-context learning, decoder-only LLMs are given 5 shots as examples. The input length is set to 1024. The details are presented in Table 6. All of our experiments were run on an A100 machine with 2 40GB GPUs.

E Related Work

We briefly overview the previous work of AQA, as well as curriculum learning. The latter is used as the baseline training approach in our study.

E.1 Attributed Question Answering

The recent studies of AQA concentrate on the exploration and application of reliable evidence from multiple sources.

ALCE (Gao et al., 2023) retrieves question-related text spans from sources. On this basis, it prompts LLMs to generate the synthetic evidence conditioned on these spans. To ensure the quality of such evidence, Gao et al. (2023) construct a benchmark dataset for fine-tuning and evaluating LLMs. Berchansky et al. (2024) propose a tailored CoT reasoning approach to assist LLMs in producing reliable evidence given miscellaneous clues. To fine-tune LLMs well within the CoT framework, Berchansky et al. (2024) introduce hybrid data into the training process, including the evidence that couples with the qualified, human-written and semi-extractive answers. All the data can be assessed from Schuster et al. (2024) 's validation set. Slobodkin et al. (2024) propose a local RAG approach to enhance AQA. This RAG method retrieves possible evidence from the sources, and uses it as constraints to the vocabulary-based answer generation.

E.2 Curriculum Learning

Curriculum Learning (CL) is an easy-to-hard training strategy in machine learning, which is motivated by human behavior (Bengio et al., 2009). Two fundamental components of CL are difficulty metrics and course selection strategy (Yuan et al., 2023).

Lao et al. (2021) incorporate the degree of visual modality dependence into the difficulty metric to mitigate training data bias. CLIMB (Martinez et al., 2023) utilizes the confusion of the current model as a dynamic difficulty score, emulating the order in which children learn languages. LLMCL-GEC (Fang et al., 2025) categorizes data into easy, medium, and difficult score bands based on LLM-assigned scores. Xu et al. (2020) also defines the difficulty through the synthesized loss of LLMs rather than relying on human intuitions. In the aspect of course selection, many approaches adopt the EXP3 algorithm (Bejan et al., 2023; Matiisen et al., 2019). To ensure a steadier transition, a

Task instruction:

You are an expert text generator required to process documents for question-answering difficulty enhancement. Given a question, useful passages, and corresponding short answers. The goal is to generate a noisy passage that distracts the model. Follow these strict guidelines:

- Generate a source that is highly topical in relevance but lacks answer-specific information.

- Ensure the generated passage cannot be used to truly answer the question.

- Output the new passage without any explanations. Don't extend 100 words!

Input:

QUESTION: When does the new army uniform come out?

SOURCE 1: First unveiled in June 2004, it is the successor to the Battle Dress Uniform (BDU) and Desert Camouflage Uniform (DCU) worn from the 1980s and 1990s through to the mid-2000s, respectively. It is also the successor to the Airman Battle Uniform for the U.S. Air Force. In early 2004, some U.S. Army soldiers in Iraq were issued the "Close Combat Uniform" a variant of the Desert Camouflage Uniform (DCU) that featured new features such as shoulder pockets affixed with hook-and-loop "Velcro" fasteners, chest-worn rank insignia, and a new collar.

SOURCE 2: After a 60-year hiatus, on November 11, 2018, the U.S. Army announced it would adopt a new uniform patterned on the "pinks and greens" effective 2020, with phase-in to be complete by 2028. The decision to adopt the new uniform was done to fill the gap between the formal blue Army Service Uniform and the Army Combat Uniform created by the deauthorization of the Green Service Uniform used between 1954 and 2015. With the change, the blue Army Service Uniform will again be restricted to ceremonial wear.

SHORT ANSWERS: June 2004; 2020

Output:

The U.S. Army has been continually updating its uniform standards to reflect modern needs and operational environments. These changes often involve integrating new fabrics and designs for improved comfort and functionality. For instance, the Army has also considered various environmental factors such as the effects of different climates on soldiers' performance and well-being, influencing uniform choices in the past few years. Moreover, the integration of more adaptive camouflage patterns continues to evolve as part of ongoing trials and testing, aimed at enhancing field effectiveness.

Table 7: An example illustrating the prompt used for constructing distracting sources. The table presents the original question, sources, and short answer as inputs, along with the corresponding generated distracting source as the output.

Task instruction:

Given a question and a relevant passage, determine if the passage contains the correct answer. If it contains, extract the answer span. Otherwise, output "no answer".

Table 8: Prompt used for validating generated sources.

Backbone	Model	R-L	F1	Rec	Avg
Encoder-decoder framework					
BART-large	+SEMQA	52.5	64.3	73.3	58.1
	+TPCL (Full)	53.7	66.7	76.4	59.8
T5-large	+SEMQA	53.1	65.1	69.8	58.8
	+TPCL (Full)	**55.3**	**65.8**	**76.7**	**60.3**
Decoder-only framework					
LLaMA-3.1	+SEMQA	38.4	43.9	54.2	41.1
	+TPCL (One-off)	45.2	50.6	57.4	47.8
Gemma-3.0	+SEMQA	49.4	63.7	77.3	56.1
	+TPCL (One-off)	**52.4**	**69.1**	**78.0**	**60.2**

Table 9: Performance (%) on the challenge test set.

linear decay formula to update the threshold over the course of training is introduced (Nguyen et al., 2025; Vakil and Amiri, 2023). CLFFRD (Xu et al., 2024) designs tri-level difficulty metrics and employs a linear pacing schedule that gradually expands the training subset. Motivated by these approaches, we propose a hybrid strategy that combines EXP3 with linear decay to achieve progressive sample selection.

F Challenge Test Set

Table 7 and Table 8 present the detailed prompts used to construct and verify the distracting sources, respectively. Table 9 shows the results obtained on the challenge test set.