

# Tackling Distractor Documents in Multi-Hop QA with Reinforcement and Curriculum Learning

Jerry Huang<sup>1\*</sup>, Siddarth Madala<sup>1</sup>, Risham Sidhu<sup>1</sup>, Cheng Niu<sup>2</sup>,  
Hao Peng<sup>1</sup>, Julia Hockenmaier<sup>1</sup>, Tong Zhang<sup>1</sup>

<sup>1</sup>University of Illinois Urbana-Champaign, <sup>2</sup>NewsBreak  
{jerry8, smadala2, rsidhu3, haopeng, juliahmr, tozhang}@illinois.edu  
cheng.niu@newsbreak.com

## Abstract

Retrieval-augmented generation (RAG) systems rely on retrieval models for identifying relevant contexts and answer generation models for utilizing those contexts. However, retrievers exhibit imperfect recall and precision, limiting downstream performance. We introduce RAG-RL, an answer generation model trained for multi-hop question answering (MHQA) to not only generate answers but also to identify and cite relevant information from larger sets of retrieved contexts, shifting some of the burden of identifying relevant documents from the retriever to the answer generator. Our approach uses *curriculum learning*, where models are trained across retrieval settings with varying levels of noise. Our experiments show that training samples with fewer distractor documents enable models to acquire citation and reasoning skills with greater sample efficiency and generalizability, demonstrating strong model performance even as the number of irrelevant passages increases. We benchmark our methods on three open-domain MHQA datasets and report significant gains in answer and citation accuracy. Furthermore, our experiments provide empirical insights into how simpler training samples can give models stronger signals for learning specific skills (e.g., citation generation) and how different components of post-training (e.g., training set construction, rule-based rewards, training sample ordering, etc.) impact final model performance.<sup>1</sup>

## 1 Introduction

Retrieval-augmented generation (RAG; Guu et al., 2020; Lewis et al., 2020; Wang et al., 2024a) relies on retrieval and generation models working together to retrieve and integrate external contexts effectively for answering questions or generating

<sup>\*</sup>This work was completed during an internship at NewsBreak.

<sup>1</sup>Code is available at <https://github.com/jh603/mhqa-cl>

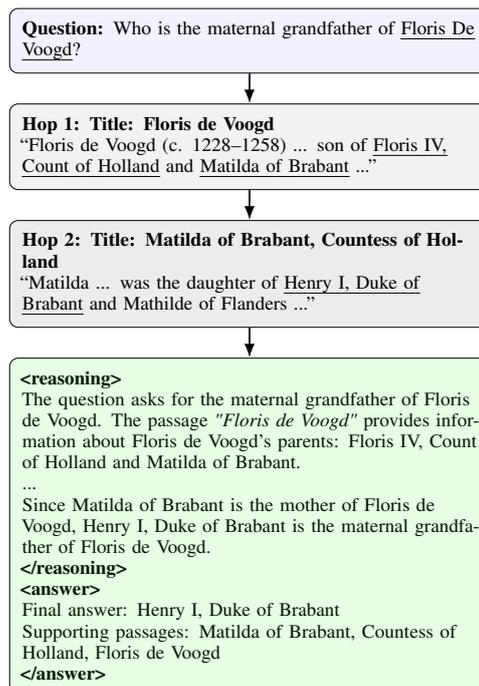


Figure 1: An example of a multi-hop reasoning chain taken from the MuSiQue dataset. RAG-RL generated the reasoning trace and final answer/citations observed in the green block.

content. While previous works have made significant progress in improving these systems by optimizing retrieval and reranking models (Zhang et al., 2024a; Gutiérrez et al., 2025; Weller et al., 2025), challenges persist when it comes to retrieving relevant real-world contexts that require reasoning, especially those whose relevance goes beyond semantic similarity (Su et al., 2024). Moreover, prior work has demonstrated that generative models often struggle to effectively integrate information across multiple documents, a limitation attributed to their constrained reasoning capabilities, particularly in domains such as code generation (Wang et al., 2024b) and in settings involving long-context retrieval (Zhou et al., 2025; Yen et al., 2025).

In this work, we tackle the aforementioned

challenges by training reasoning language models (RLMs) capable of performing reasoning over a greater number of retrieved documents. Prior approaches for improving RAG have focused on optimizing the retrieval and reranking components by maximizing metrics such as recall@5. In contrast, we propose shifting some of the retrieval burden from the retriever to the generation model itself. An answer generation model that can effectively differentiate between relevant and irrelevant contexts, when given a longer list of retrieved passages, would reduce dependence on high-precision retrieval and increase recall by instead maximizing metrics such as recall@10 or recall@25 (Jin et al., 2025).

Building on the recent success of reinforcement learning (RL) in enhancing the reasoning capabilities of LLMs in the domains of mathematics and coding (Wei et al., 2025; Xie et al., 2025), we apply Group Relative Policy Optimization (GRPO; Shao et al., 2024) with simple, rule-based rewards to improve the answer generation component of RAG systems. We show that RAG-RL achieves substantial performance gains in both answer and citation generation on three open-domain multi-hop question-answer datasets. Furthermore, our comprehensive evaluation demonstrates that RAG-RL achieves improved performance both in distractor-rich and gold-only settings.<sup>2</sup> These settings respectively mirror the use of a weaker retrieval model or a more advanced retrieval and/or reranking systems, and demonstrate that RAG-RL can be used in conjunction with past works on improving retrieval models for further improved performance.

We also conduct a comprehensive study on how different curriculum learning settings affect model performance in post-training. Specifically, we study the effectiveness of introducing question-answer training samples of varying difficulty levels and the impact that the ordering of the training set has on the final performance of the model. We observe that (1) adding easier samples during training teaches the model to more quickly learn how to generate citations as the model no longer has to identify which contexts are relevant, (2) curricula that scale problem difficulty linearly from easiest to hardest perform worse when compared to min-max curricula that begin with the easiest samples and jump straight to the hardest samples,

---

<sup>2</sup>Gold documents are documents from which the answer to a given question can be deduced, while distractor documents are those that do not contain relevant information.

and (3) the benefits of deliberately ordering training samples from easiest to hardest, as proposed by previous curriculum learning studies (Bengio et al., 2009) are not conclusively supported in RL-based post-training. *These empirical observations suggest that constructing training sets of different difficulty levels can increase sample efficiency and generalization by targeting specific skills.*

In summary, the main contributions of this work are as follows:

- We introduce RAG-RL, a model specifically trained for MHQA answer generation in RAG, using RL and curriculum learning.
- We train models across a comprehensive set of different curriculum construction and curriculum learning settings.
- We provide several empirical insights on the effectiveness of different curricula, and how different aspects of the post-training process contribute to final model performance.

## 2 Related Work

### 2.1 RAG Systems

Rather than relying solely on parametric knowledge, RAG has been widely used in tasks that require external information (Guu et al., 2020; Lewis et al., 2020; Wang et al., 2024a). Previous works have made tremendous progress in designing and training sophisticated retrieval and reranking models (Gutiérrez et al., 2025; Weller et al., 2025) for open-domain question answering (Chen et al., 2017). One important line of work has focused on improving the encoder models that are used in the embedding generation process (Lee et al., 2025a; Muennighoff et al., 2025), while another has focused on designing retrieval systems that focus on drawing connections between multiple different documents (Guo et al., 2024; Gutiérrez et al., 2025). Rank1 (Weller et al., 2025) has also recently demonstrated that allocating test-time compute for document reranking can lead to performance improvements when retrieving contexts that require in-depth reasoning. Past work has also sought to take advantage of the long context lengths of modern-day LLMs by providing these models with larger sets of retrieved documents, but have shown that these models struggle to effectively identify relevant contexts as the number of retrieved passages increases (Jin et al., 2025; Zhou et al., 2025; Yen et al., 2025).

## 2.2 Multi-Hop Question Answering

A multi-hop question requires combining information across multiple passages and performing reasoning to arrive at a correct answer (Mavi et al., 2024; Nishida et al., 2019). The number of pieces of information required to successfully answer the question is referred to as the number of hops. The terms passages and documents are used interchangeably to denote disjoint contexts that are retrieved by a retrieval model. Figure 1 demonstrates how RAG-RL operates in the multi-hop question-answering setting.

## 2.3 Reasoning Language Models

With the introduction of RLMs in OpenAI’s o1 models (OpenAI et al., 2024), the research community has made progress in replicating similar models that have shown impressive performance in tasks that require reasoning, driven in part due to R1’s release (DeepSeek-AI et al., 2025). Prior works have demonstrated the potential for training smaller-scale RLMs in the domains of mathematics, logic, and coding (Xie et al., 2025; Wei et al., 2025) and have also achieved impressive performance. However, to the best of our knowledge, no one has trained RLMs specifically for the answer generation component of RAG.

## 2.4 Curriculum Learning

Curriculum learning (Bengio et al., 2009) has been extensively studied as a training paradigm that orders training samples by increasing difficulty, leading to improved generalization. In question answering (QA), it has been used to reduce distributional shifts between pre-training and downstream fine-tuning datasets (Zhang et al., 2024b). Recent advances in LLMs have incorporated curriculum-inspired self-improvement mechanisms (Lee et al., 2025b), where models iteratively augment their training data with instances they can already solve, to facilitate generalization to slightly more complex reasoning tasks. In RL, curriculum learning has also been applied to gradually expose agents to more challenging environments (Narvekar et al., 2020); however, its effectiveness remains task-dependent, with some studies reporting only marginal gains (Xie et al., 2025).

## 3 RAG-RL

In this section, we include a detailed overview of the training process for RAG-RL. We outline the

rule-based rewards used in the policy update algorithm and then introduce the curriculum construction settings used in our experiments.

### 3.1 Reward Modeling

In our work, we use RL as our post-training method as it eliminates the need for training sets consisting of high-quality supervised trajectories produced either by humans or stronger models.

Our rule-based rewards consist of three components: answer rewards, citation rewards, and formatting rewards. Importantly, we note that to receive any answer or citation rewards, proper formatting is a prerequisite.

**Answer Rewards** To incentivize correct final answers, we define the answer reward as:

$$\mathcal{R}_{\text{answer}} = \gamma_{\text{answer}} \cdot 1(o_{\text{answer}} = G_{\text{answer}}), \quad (1)$$

where  $o_{\text{answer}}$  is the generated final answer,  $G_{\text{answer}}$  is the ground truth answer, and  $\gamma_{\text{answer}}$  is a scaling factor, which we set to 5 for our experiments.

**Citation Rewards** To reward correct citations, we define the citation reward as:

$$\mathcal{R}_{\text{citations}} = \gamma_{\text{correct}} \cdot \text{Recall}(o_{\text{citations}}, G_{\text{citations}}) - \gamma_{\text{incorrect}} \cdot C_{\text{incorrect}} \quad (2)$$

where recall denotes the fraction of relevant citations cited in the final answer  $o_{\text{citations}}$ ,  $G_{\text{citations}}$  is the list of ground truth citations,  $C_{\text{incorrect}}$  is the number of incorrect citations, and both  $\gamma_{\text{correct}}$  and  $\gamma_{\text{incorrect}}$  are the scaling factors which we set to 5 and 2 respectively.

**Formatting Rewards** To enforce the desired output format, we assign a reward of  $\gamma_{\text{format}}$  for correct formatting (i.e., the presence of proper XML tags and required headings) while imposing a penalty  $p$  for outputs with excessive text or non-English Unicode characters.<sup>3</sup> Formally, we define the reward as:

$$\mathcal{R}_{\text{formatting}} = \begin{cases} \gamma_{\text{format}}, & \text{if formatting is correct} \\ p, & \text{otherwise.} \end{cases} \quad (3)$$

<sup>3</sup>This penalty has proven particularly beneficial for improving training stability by encouraging the model to generate responses in English.

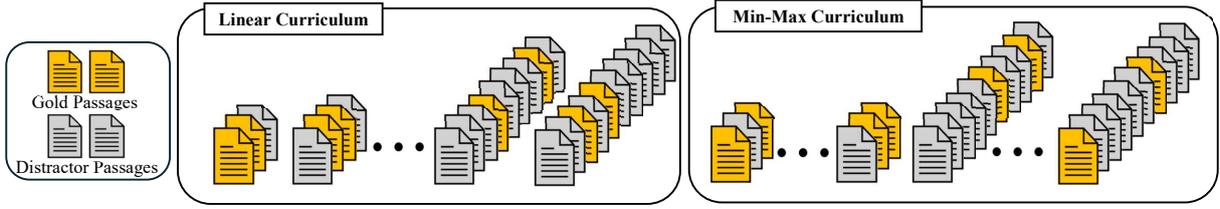


Figure 2: Overview of two curriculum construction settings used during training. Linear denotes a curriculum that scales the difficulty level (the number of distractor passages) from 1 to  $H$ , while min-max denotes a curriculum that is split evenly between the easiest and the hardest problems.

**Total Reward and Objective Function** The overall reward for a training sample is the sum of the individual components:

$$\mathcal{R}_{\text{total}} = \mathcal{R}_{\text{answer}} + \mathcal{R}_{\text{citation}} + \mathcal{R}_{\text{formatting}}.$$

This reward is then used in the GRPO algorithm for policy optimization (Shao et al., 2024). The scaling constants we choose in our experiments weigh correctness more than formatting, and in preliminary experiments, we observed no significant changes in performance when adjusting these parameters by small amounts.

### 3.2 Curriculum Construction

Curriculum construction builds training sets by selecting or generating samples across a range of difficulty levels. In this work, we investigate two main difficulty axes: (1) synthetic difficulty, where training samples are algorithmically constructed to span predefined difficulty levels, and (2) accuracy-based difficulty, where the base model’s performance determines which samples it can or cannot solve before any post-training is applied.

**Synthetic Difficulty** Given a question  $Q$ , and a set of documents  $D$ , we partition  $D$  into a set of gold documents  $D^+$  and a set of distractor documents  $D^-$ . The number of hops required to correctly answer  $Q$  is given by  $j = |D^+|$ , while  $h = |D^-|$  represents the number of retrieved distractor documents. Naturally, the difficulty of a multi-hop question can be measured along two dimensions: the number of hops required  $j$  and the number of distractor documents provided to the generation model from  $D^-$ . For training RAG-RL, we define the difficulty of a training sample solely based on the subset size of  $D^-$  that we provide to the generation model. The easiest training samples contain at most one distractor document along with all gold documents, while the hardest samples

include the full set of all retrieved documents. Formally, a training sample  $S_i$  of difficulty level  $l$  is defined as

$$S_i^l = [Q, \{D_1^+, D_2^+, \dots, D_j^+, D_1^-, D_2^-, \dots, D_d^-\}],$$

where  $d = \min(\max(l + 2 - j, 0), h)$ ,

where the order of the documents in each  $S_i$  is shuffled to ensure a realistic retrieval setting.

Since the minimum number of hops required among all of our datasets is 2, a difficulty level of 1 corresponds to 1 distractor document for a 2-hop question. It follows that the highest difficulty level we can effectively introduce is thus  $h$ , which we denote as  $H$  going forward. This definition of synthetic difficulty ensures that all gold contexts are retrieved regardless of difficulty level.

While the datasets we use contain 2-hop, 3-hop, and 4-hop questions, we focus on the number of distractor documents as the primary axis of difficulty. This choice is motivated by the limited granularity offered by hop count alone, as the vast majority of questions in all three datasets we use are 2-hop questions. For completeness, we include an ablation in Section 5.4 that jointly considers both the number of hops and distractor documents by sorting each question by the number of hops and then augmenting each question to span a pre-defined curriculum. Moreover, in Appendix C.5 we present results that show a negative correlation between model performance and the number of hops in each question.

**Accuracy-Based Difficulty** An alternative way to define the difficulty of a training sample is to benchmark the base model’s performance on each sample. Specifically, we compute the pass@k of each training sample and partition the dataset into two subsets: samples with pass@k = 1 and samples with pass@k = 0. We refer to the former as *base-answerable* and the latter as *base-unanswerable*.

A prediction is considered correct if the generated final answer achieves an F1 score of 1 when compared to the ground truth answer.

## 4 Experiments

### 4.1 Datasets

We evaluate RAG-RL on three open-domain multi-hop question answering benchmarks: **HotpotQA** (Yang et al., 2018), **MuSiQue** (answerable) (Trivedi et al., 2022), and **2Wiki** (Ho et al., 2020). While HotpotQA has been shown to be a weaker test for multi-hop reasoning due to the presence of spurious signals (Trivedi et al., 2023), we include it due to its widespread use but mainly focus on the other two datasets in our discussions.

### 4.2 Training Setup

We use Qwen2.5-7B-Instruct (Qwen et al., 2025) as our base model and employ GRPO for the post-training process.<sup>4</sup> Approximately five thousand QA pairs are sourced from each dataset’s respective training sets. All experiments train for a single epoch with a constant learning rate of 1.0e-6, a global batch size of 294, KL coefficient of 0.01, and 7 rollouts for each of the 42 problems in each batch. Each model was trained on Nvidia H100 GPUs. While we observe that reward signals continue to improve beyond 120 steps, we limit training to 5,000 samples due to computational constraints. To assess the generalizability of our experiments to larger training sets, we include additional runs in Appendix C.1, which confirm that *the trends and insights we report hold consistently at scale*. Additional dataset construction and training details can be found in Appendices A and B.

### 4.3 Baselines

We report the performance of our base model, Qwen2.5-7B-Instruct, as our first baseline. For an additional fine-tuned baseline (FT-Baseline), we use the max curriculum, which fine-tunes the model on samples at the highest difficulty level  $H$ , matching the difficulty of problems expected at test-time. To compare our methods with previous works, we measure the performance of our models in two settings: the *distractor setting* and the *ideal retrieval setting*. The distractor setting consists of providing the generation model all gold passages and up

<sup>4</sup>Additional experimental results using Llama3.1-8b-Instruct and supervised fine-tuning instead of RL can be found in Appendices C.2 and C.3.

to 18 distractor passages, which is comparable to having the reasoning model handle both reranking and answer generation. On the other hand, in the ideal retrieval setting, the reasoning model is given only the gold truth passages, which is comparable to using a strong retrieval and reranking system.

Previous works on improving multi-hop question-answer performance, such as Beam Retrieval (Zhang et al., 2024a) and Smoothing R3 (Yin et al., 2023), have primarily focused on optimizing the retrieval component of RAG and utilize span prediction models for answer generation, thus making a direct comparison of generator performance difficult. To better isolate and evaluate generation quality, we adopt the ideal retrieval setting as a more controlled and comparable benchmark. When comparing the performance of RAG-RL to the few past studies that have focused on improving answer generation models (Jin et al., 2025; Zhang et al., 2024c), RAG-RL achieves SOTA performance.

### 4.4 Curriculum Learning Settings

To investigate the effectiveness of curriculum construction in the post-training process, we benchmark several different curricula. As defined in Section 3.2, synthetic difficulty levels range from 1 to  $H$ , while accuracy-based difficulty partitions training samples into base-answerable and base-unanswerable subsets. Figure 2 provides an illustration of the main synthetic curricula used in our experiments. We define a function  $C_{\text{setting}} : \{1, \dots, n\} \rightarrow \{1, \dots, H\}$  that maps an index  $i$  in the training set to its corresponding difficulty level under each setting. We set  $H$  to be 10, 20, and 10 for HotpotQA, MuSiQue, and 2Wiki respectively.

#### 4.4.1 Synthetic Curricula Variants

- **Max:** Each sample in the training set is presented at the maximum difficulty level (i.e., the difficulty level expected at test time). Thus, the difficulty function is defined as:

$$C_{\text{max}}(i) = H, \quad \forall i \in \{1, \dots, n\}$$

- **Linear:** The training set is partitioned into  $K$  equally sized subsets, with difficulty levels increasing linearly from 1 to  $K$ . The mapping function is thus:

$$C_{\text{linear}}(i) = \left\lceil \frac{H \cdot i}{n} \right\rceil$$

Eval. Setting	Curriculum	HotpotQA			MuSiQue			2Wiki		
		Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1
Distractor	Baseline	60.65	36.47	45.55	25.88	25.35	25.61	48.71	40.18	44.03
	FT-Baseline (Max)	66.04	73.93	69.76	40.91	53.07	46.20	67.99	77.08	72.25
	Linear	68.71	78.54	73.30	44.68	59.79	51.14	68.92	<b>83.23</b>	<b>75.40</b>
	Min-Max	<b>68.87</b>	<b>81.64</b>	<b>74.72</b>	<b>47.18</b>	<b>64.48</b>	<b>54.49</b>	<b>70.77</b>	76.37	73.46
	Base-Answerable	66.19	72.81	69.34	41.40	54.28	46.97	68.56	76.36	72.25
	Base-Unanswerable	65.25	71.13	68.06	38.84	52.15	44.52	66.59	72.04	69.21
Ideal Retrieval	Baseline	67.90	63.26	65.50	41.16	58.16	48.21	70.29	54.49	61.39
	FT-Baseline (Max)	74.25	86.26	79.81	54.64	68.84	60.92	71.82	78.53	75.02
	Linear	75.67	89.34	81.94	61.10	73.90	66.89	74.31	<b>88.82</b>	<b>80.92</b>
	Min-Max	<b>76.18</b>	<b>93.13</b>	<b>83.81</b>	<b>65.06</b>	<b>81.51</b>	<b>72.37</b>	<b>75.06</b>	80.37	77.63
	Base-Answerable	74.76	83.69	78.98	57.53	67.75	62.22	72.61	79.98	76.11
	Base-Unanswerable	75.23	82.99	78.92	54.53	68.37	60.67	71.04	74.30	72.63

Table 1: Model performance under the distractor and ideal retrieval evaluation settings across different curriculum construction settings. We use up to 5,000 training samples for all runs as outlined in Section 4.2. The best-performing curriculum for each metric is bolded. Additional training runs with larger training sets are provided in Appendix C.1.

- **Min-Max:** The training set is split into two equal parts, where the first half consists of the easiest difficulty level (1) and the second half consists of the hardest difficulty level ( $H$ ). The function is defined as:

$$C_{\text{min-max}}(i) = \begin{cases} 1, & \text{if } i \leq n/2 \\ H, & \text{if } i > n/2 \end{cases}$$

#### 4.4.2 Accuracy-Based Curricula

- **Base-Answerable:** The training set includes only samples from the maximum synthetic difficulty level that are base-answerable.
- **Base-Unanswerable:** The training set includes only samples from the maximum synthetic difficulty level that are base-unanswerable.

#### 4.5 Evaluation

To benchmark the performance of our RLMs, we evaluate the F1 scores of the generated answer and passage-level citations on the validation sets provided by our selected benchmark datasets. We sample each response 3 times and take the average F1 score among all generations. Joint F1, which captures both answer and citation correctness, serves as our primary metric. Dataset statistics can be found in Appendix A.

### 5 Results

Our results section is organized into subsections, beginning with the main findings, followed by an investigation of several research questions. The primary goals of our experiments are twofold: first,

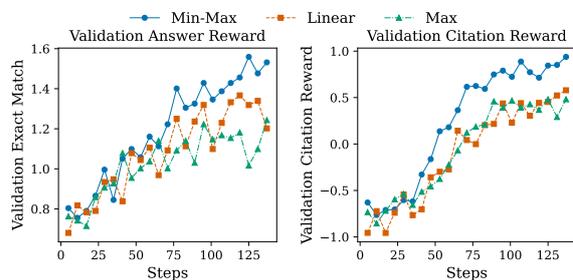


Figure 3: Plots of validation answer and citation rewards during training for three curricula on the MuSiQue dataset.

Metric	Answer F1	Citation F1	Joint F1
Pass@1	25.42	25.11	25.26
Pass@32	43.38	45.46	44.40

Table 2: Base model pass@k on the MuSiQue dataset.

to evaluate whether curriculum learning can enhance RL-based post-training; and second, to identify which components of the post-training pipeline contribute most significantly to final model performance and why.

#### 5.1 Main Results

Table 1 presents the performance of our baseline model and the RLMs we trained under our outlined curriculum construction settings in the distractor and ideal retrieval evaluation settings. The results strongly support the notion that curriculum learning can help improve RL-based post-training. Across all three datasets and both evaluation settings, the min-max and linear curricula achieve the highest joint F1, improving over the max curriculum by a margin of 3 to 8 points.

Eval. Setting	Sample Ordering	Answer F1	Citation F1	Joint F1
Distractor	Sorted by Difficulty	44.68	59.79	51.14
	Randomly Shuffled	44.27	60.35	51.07
	Sorted by Hops and Difficulty	43.78	61.20	51.04
Ideal	Sorted by Difficulty	61.10	73.90	66.89
	Randomly Shuffled	60.80	73.47	66.53
	Sorted by Hops and Difficulty	60.94	75.16	67.31

Table 3: Ablation study on training sample ordering using the linear curriculum on the MuSiQue dataset. Results are reported under both distractor and ideal retrieval settings.

## 5.2 Does adding easier samples improve performance?

The primary difference between our synthetic curricula and the max variant is the introduction of easier training samples. The results in Table 1 show that all the curricula that contain easier samples outperform the max curriculum. Figure 3 shows that the validation citation reward climbs much faster for the min-max curriculum, suggesting that training samples with few distractor passages allow the model to learn how to generate correct citations with greater sample efficiency. Given a large number of distractor passages, the model must first identify potential candidates and then reason over them. These findings indicate that easier samples in the synthetic curricula serve to teach citation skills early in training, which can then be refined by harder examples that require multi-step reasoning over longer contexts.

## 5.3 Do we need granular problem difficulty?

Previous work in the area of self-improvement has shown that LLMs exhibit limited generalizability and that gradually increasing the difficulty levels of training samples from weak-to-strong is effective for helping models generalize beyond their initial training distributions (Lee et al., 2025b). However, our results suggest that this is not always necessary as the min-max curriculum outperforms the linear curriculum in most cases. We believe this to be a byproduct of our base model having relatively strong performance on the task before any post-training is applied. We show our base model’s pass@32 in Table 2 and find that these scores are comparable to our fine-tuned baseline.

## 5.4 Does dataset ordering matter?

Curriculum learning strategies where training sets are ordered from easy to hard have been successfully employed in many areas of machine learning (Soviany et al., 2022). To assess the impact

of sample ordering on model performance, we experiment with three variants of a linear curriculum: (i) samples sorted by difficulty, measured as the number of distractors; (ii) samples presented in a randomly shuffled order; and (iii) samples first ordered by the number of hops and then augmented to span the linear curriculum. As shown in Table 3, the resulting F1 scores across these curricula are broadly comparable. These findings suggest that, in this setting, the specific ordering of training examples does not yield consistent or significant performance differences.

## 5.5 Training on base-answerable versus base-unanswerable samples

We observe that models trained on base-answerable samples consistently outperform those trained on base-unanswerable samples. As shown in Table 4, this trend holds across both the distractor and ideal retrieval settings. We hypothesize that the superior performance from training on base-answerable samples stems from their alignment with the base model’s pretrained capabilities. This is reminiscent of findings from Zhang et al. (2025), which demonstrate that supervised fine-tuning is most effective when responses are aligned with the base model’s pretrained distribution.

Another possibility for the worse performance from base-unanswerable samples stems from these samples producing a greater number of groups where the rewards among all samples are equal, resulting in a computed advantage of zero and thus no gradient signal nor policy update. The Dynamic sAmpling Policy Optimization (DAPO; Yu et al., 2025) algorithm also notes this limitation, and they address this challenge by over-sampling across the entire dataset and discarding prompts that yield uniform rewards across generated responses. Based on our results, we believe tackling this issue from a curriculum construction angle can also yield improved training efficiency.

## 5.6 What rule-based rewards matter?

Recent works have observed that performance gains from using algorithms such as GRPO might stem from improved output formatting rather than improvements in reasoning ability (Petrov et al., 2025). To isolate this phenomenon and see how answer and citation rewards contribute to final model performance, we conduct an ablation where we train our models with only formatting rewards (Eq. 3), omitting the answer (Eq. 1) and citation

Eval. Setting	Curriculum	HotpotQA			MuSiQue			2Wiki		
		Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1
Distractor	Baseline	61.31 / 9.21	44.89 / 35.65	51.83 / 14.64	40.96 / 4.63	27.91 / 19.25	33.20 / 7.46	61.64 / 6.75	42.71 / 33.80	50.45 / 11.25
	FT-Baseline (Max)	82.63 / 18.07	76.01 / 68.70	79.18 / 28.61	63.87 / 13.51	55.54 / 47.51	59.41 / 21.04	84.42 / 17.37	76.53 / 75.50	80.28 / 28.25
	Linear	86.26 / <b>20.48</b>	80.84 / 73.02	83.46 / 31.99	70.18 / 18.82	62.70 / 53.70	66.23 / 27.87	85.00 / <b>23.14</b>	<b>83.77 / 80.32</b>	<b>84.38 / 35.93</b>
	Min-Max	<b>86.27 / 20.40</b>	<b>83.21 / 76.46</b>	<b>84.71 / 32.20</b>	<b>71.73 / 22.89</b>	<b>67.42 / 58.47</b>	<b>69.51 / 32.90</b>	<b>87.83 / 20.41</b>	77.66 / 74.19	82.43 / 32.02
	Base-Answerable	82.63 / 17.09	75.08 / 66.20	78.67 / 27.16	65.62 / 16.17	57.97 / 47.65	61.56 / 24.15	86.26 / 17.42	77.29 / 75.14	81.53 / 28.28
	Base-Unanswerable	82.94 / 16.71	72.74 / 65.35	77.51 / 26.62	63.59 / 14.49	55.44 / 45.81	59.23 / 22.01	83.92 / 16.16	71.42 / 73.19	77.17 / 26.48

Table 4: Model performance under the distractor evaluation setting evaluated on [base-answerable](#) and [base-unanswerable](#) evaluation sets. We use up to 5,000 training samples for all runs as outlined in Section 4.2. Results under the ideal retrieval setting can be found in the Appendix in Table 12.

Eval. Setting	Curriculum	Answer F1	Citation F1	Joint F1
Distractor	Baseline	25.88	25.35	25.61
	FT-Baseline	40.91 / 37.97	53.07 / 36.44	46.20 / 37.19
	Min-Max	47.18 / 41.30	64.48 / 42.21	54.49 / 41.75
Ideal	Baseline	41.16	58.16	48.21
	FT-Baseline	54.64 / 50.59	68.84 / 55.58	60.92 / 52.97
	Min-Max	65.06 / 57.95	81.51 / 57.18	72.37 / 57.56

Table 5: MuSiQue formatting ablation. The first number in each cell denotes the use of both accuracy and formatting rewards, while the second number in each cell denotes the use of only formatting rewards.

rewards (Eq. 2).

Table 5 presents results on the MuSiQue dataset that demonstrate that employing formatting rewards alone using the min-max curriculum achieves higher answer F1 compared to that of the max curriculum with all rewards. While adding additional answer and citation rewards increases the sample efficiency of the post-training process, this ablation demonstrates that choosing the appropriate training curriculum also plays a major role. We propose that post-training using formatting rewards alone can serve as a stronger baseline before incorporating accuracy-based metrics.

## 5.7 Performance on previously unsolved samples

To evaluate the extent to which RL-based post-training and curriculum learning enhance model performance on previously unsolved questions, we additionally partition the evaluation set into *base-answerable* and *base-unanswerable* subsets. As shown in Table 4, both the min-max and linear curricula yield the greatest gains on the base-unanswerable subset. Nevertheless, a notable performance gap remains between the two subsets. These results suggest that, while curriculum learning can improve a model’s ability to address previously challenging examples, generalization to out-of-distribution samples remains a key limitation of RL-based post-training (Xiong et al., 2025a).

Curriculum	Mean Tokens	Median Tokens
Baseline	168	152
FT-Baseline (Max)	185	160
Linear	203	168
Min-Max	220	179

Table 6: Response length statistics on the MuSiQue dataset under the distractor setting. We report the mean and median number of generated tokens for each training curriculum.

## 5.8 Response Length Across Curricula

Table 6 reports statistics on the average completion length of each our models, measured as the number of tokens generated per answer for each training curriculum on the MuSiQue dataset. Compared to the baseline model, all curriculum-trained models exhibit a moderate increase in response length. We observe a correlation between the mean number of generated tokens and overall curriculum performance during post-training.

## 6 Discussion

In this work, we introduce RAG-RL, a reasoning language model specifically trained for the answer generation component of RAG for MHQA. Our experiments demonstrate that stronger answer generation models can reduce the burden on retrieval models by reasoning over larger sets of retrieved contexts and that curriculum learning is a powerful tool for improving sample efficiency and generalization during post-training.

Our experiments and ablations support the following key observations: (1) curriculum construction is a powerful method for improving post-training performance, (2) easier training samples (i.e., those with fewer distractor documents) provide models with a stronger signal for learning how to generate citations, and (3) LLMs do not necessarily benefit the most from curricula with gradually increasing difficulty levels (i.e., min-max

performs better than a linear curriculum in most of our experiments). While RL-based post-training methods have unlocked a new dimension of scaling for LLMs, our experiments take an in-depth look into understanding which components of these post-training methods contribute to improving model performance and why.

**Future Work** Our findings suggest that LLMs exhibit limited generalization with performance gains extending only marginally beyond the training distribution. To address this, we propose a systematic categorization of training samples into base-answerable and base-unanswerable instances. Empirically, we find that incorporating synthetic difficulty levels in curriculum construction enables models to acquire citation and reasoning skills with greater sample efficiency and generalizability. Exploring algorithmic curriculum generation methods that target specific areas of improvement, especially for tasks lacking natural difficulty levels, is an exciting research direction.

## 7 Limitations

While RAG-RL achieves strong performance across multiple multi-hop QA benchmarks, a key limitation of our experimental setup is the assumption that all relevant gold documents are present in the retrieved set. In real-world scenarios, this assumption may not hold. We suggest explicitly training the model to return an “unanswerable” response when the retrieved context lacks sufficient information. Additionally, our curriculum construction process employs a static progression through difficulty levels, advancing to harder samples regardless of whether the model has fully converged on earlier stages. In contrast, prior work on curriculum learning often adopts adaptive schedulers that revisit easier examples and only proceeds to more difficult ones once the model’s performance plateaus. Incorporating such adaptive strategies may further enhance training efficiency and generalization, and we leave this as an avenue for future exploration.

## Acknowledgments

## References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *International Conference on Machine Learning*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine

Bordes. 2017. [Reading Wikipedia to answer open-domain questions](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and 1 others. 2025. [Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. [Raft: Reward ranked finetuning for generative foundation model alignment](#). *arXiv preprint arXiv:2304.06767*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv e-prints*, pages arXiv–2407.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. [Lightrag: Simple and fast retrieval-augmented generation](#). *Preprint*, arXiv:2410.05779.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. [From rag to memory: Non-parametric continual learning for large language models](#). *Preprint*, arXiv:2502.14802.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. [Realm: retrieval-augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2025. [Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG](#). In *The Thirteenth International Conference on Learning Representations*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025a. [Nv-embed: Improved techniques for training llms as generalist embedding models](#). *Preprint*, arXiv:2405.17428.

Nayoung Lee, Ziyang Cai, Avi Schwarzschild, Kangwook Lee, and Dimitris Papailiopoulos. 2025b. [Self-improving transformers overcome easy-to-hard and length generalization challenges](#). *Preprint*, arXiv:2502.01612.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. [Multi-hop question answering](#). *Found. Trends Inf. Retr.*, 17(5):457–586.
- Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. [Generative representational instruction tuning](#). *Preprint*, arXiv:2402.09906.
- Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. 2020. [Curriculum learning for reinforcement learning domains: A framework and survey](#). *Preprint*, arXiv:2003.04960.
- Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. [Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2335–2345, Florence, Italy. Association for Computational Linguistics.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, and 1 others. 2024. [Openai ol system card](#). *Preprint*, arXiv:2412.16720.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. 2025. [Proof or bluff? evaluating llms on 2025 usa math olympiad](#). *Preprint*, arXiv:2503.21934.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and 1 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. [Curriculum learning: A survey](#). *Preprint*, arXiv:2101.10382.
- Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Serkan O. Arik, Danqi Chen, and Tao Yu. 2024. [Bright: A realistic and challenging benchmark for reasoning-intensive retrieval](#). *Preprint*, arXiv:2407.12883.
- H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal. 2022. [Musique: Multihop questions via single-hop question composition](#). *Transactions of the Association for Computational Linguistics*, 10:539–554.
- H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal. 2023. [Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.
- Yuhao Wang, Ruiyang Ren, Junyi Li, Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024a. [REAR: A relevance-aware retrieval-augmented framework for open-domain question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5613–5626, Miami, Florida, USA. Association for Computational Linguistics.
- Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F. Xu, Yiqing Xie, Graham Neubig, and Daniel Fried. 2024b. [Coderag-bench: Can retrieval augment code generation?](#) *Preprint*, arXiv:2406.14497.
- Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I. Wang. 2025. [Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution](#). *Preprint*, arXiv:2502.18449.
- Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. [Rank1: Test-time compute for reranking in information retrieval](#). *Preprint*, arXiv:2502.18418.
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. [Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning](#). *Preprint*, arXiv:2502.14768.
- Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khalman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, Chi Jin, Tong Zhang, and Tianqi Liu. 2025a. [Building math agents with multi-turn iterative preference learning](#). In *The Thirteenth International Conference on Learning Representations*.
- Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, and 1 others. 2025b. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*.

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](#). *Preprint*, arXiv:1809.09600.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. [HELMET: How to evaluate long-context models effectively and thoroughly](#). In *The Thirteenth International Conference on Learning Representations*.

Zhangyue Yin, Yuxin Wang, Xiannian Hu, Yiguang Wu, Hang Yan, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. 2023. [Rethinking label smoothing on multi-hop question answering](#). *Preprint*, arXiv:2212.09512.

Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. [Dapo: An open-source llm reinforcement learning system at scale](#). *Preprint*, arXiv:2503.14476.

Dylan Zhang, Qirun Dai, and Hao Peng. 2025. The best instruction-tuning data are those that fit. *arXiv preprint arXiv:2502.04194*.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. 2024a. [End-to-end beam retrieval for multi-hop question answering](#). *Preprint*, arXiv:2308.08973.

L. Zhang, Quan Wang, Benfeng Xu, Yi Liu, and Zhen-dong Mao. 2024b. [Curriculum learning driven domain adaptation for low-resource machine reading comprehension](#). *IEEE Signal Processing Letters*, 31:2650–2654.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024c. [Raft: Adapting language model to domain specific rag](#). *Preprint*, arXiv:2403.10131.

Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. 2025. [Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity?](#) *arXiv preprint arXiv:2502.05252*.

## A Dataset Construction

To construct our synthetic difficulty-based curricula, we randomly sample 5,000 training samples from each of the respective training sets provided by HotpotQA, MuSiQue, and 2Wiki. These datasets each provide distractor passages on similar topics that serve to increase the difficulty of the QA task compared to sampling random passages from their corpora.

To construct the base-answerable and base-unanswerable accuracy-based curricula, we start by randomly selecting up to 40,000 training samples from each of the respective training sets, and we use our base model to partition the questions into those with  $\text{pass}@k = 1$  for base-answerable and  $\text{pass}@k = 0$  for base-unanswerable. We take the maximum answer F1 score across 8 generations for each question to assess  $\text{pass}@k$ . We then clip the number of training samples so that both the base-answer and base-unanswerable training sets have the same number of training samples (approximately 5,000).

We provide the dataset statistics for all of our datasets in Tables 7 and 8 and note that the respective number of samples in each dataset are dependent on the base model’s performance, as well as the respective sizes of each dataset’s original training and evaluation set sizes.

Curriculum	Number of Training Samples		
	HotpotQA	MuSiQue	2Wiki
FT-Baseline (Max)	5000	5000	5000
Linear	5000	5000	5000
Min-Max	5000	5000	5000
Base-Answerable	3991	4057	4468
Base-Unanswerable	3991	4057	4468

Table 7: Number of training samples for each curriculum and dataset.

Setting	Number of Evaluation Samples		
	HotpotQA	MuSiQue	2Wiki
Distractor	1000	1000	1000
Ideal Retrieval	1000	1000	1000
Distractor (Base-Answerable/Unanswerable)	810	681	1000
Ideal Retrieval (Base-Answerable/Unanswerable)	512	379	1000

Table 8: Number of evaluation samples for each dataset and evaluation setting.

Eval. Setting	Curriculum	Answer F1	Citation F1	Joint F1
Distractor	Baseline	25.88	25.35	25.61
	Max	42.76	63.33	51.05
	Linear	44.05	63.08	51.87
	Min-Max	26.40	38.41	31.29

Table 9: Model performance when using supervised fine-tuning as the post-training algorithm.

## B Training

We use the Axolotl package<sup>5</sup> to conduct our post-training. For GRPO specifically, we use a beta of 0.01 and a clipping parameter of 0.2. A full list of all the hyperparameters used during training can be found in our code in the supplementary materials. We plan to release all of our code, data, and models as an open-source GitHub repository at the conclusion of the review process.

## C Additional Experiments

### C.1 Training on Larger Train Sets

The training runs presented in Section 5 are based on subsets of up to 5,000 samples, constrained by computational limitations and the large number of post-training runs required across our curricula. To assess the scalability of our findings to larger training set sizes, we include additional results in Table 11, using 40,000 training samples for HotpotQA and 19,900 for MuSiQue. These results confirm that the trends and conclusions in the main paper continue to hold at larger scales. We do not include experiments with base-answerable and base-unanswerable subsets in this setting due to the reduced number of samples remaining after filtering for base-unanswerable instances.

### C.2 Curriculum Learning with Supervised Fine-Tuning

While our work primarily focuses on using GRPO, we include additional experiments when using supervised fine-tuning (SFT; Ouyang et al., 2022) in Table 9. To create a high-quality training set for SFT, we use a rejection sampling approach (Dong et al., 2023) by generating 8 responses per question and using only the correct responses for fine-tuning. The results in Table 9 show that SFT outperforms GRPO in the max curriculum setting, demonstrating the importance of data quality and the strong performance of rejection sampling that has also been noted in Xiong et al., 2025b. Additionally, we

<sup>5</sup><https://github.com/axolotl-ai-cloud/axolotl>

Eval. Setting	Curriculum	Answer F1	Citation F1	Joint F1
Distractor	Baseline	28.04	31.47	29.65
	FT-Baseline	45.29	66.86	54.00
	Linear	42.70	61.36	50.35
	Min-Max	48.66	68.70	56.97

Table 10: Model performance on MuSiQue when using *Llama-3.1-8B-Instruct* as the base model.

see that the min-max curriculum performed quite poorly when compared to the max and linear curricula. We found that this stemmed from the fact that it was quite difficult to generate a large number of high-quality training samples at the highest difficulty level.

### C.3 Using Llama as a Base Model

We include a subset of our experiments where *Llama-3.1-8B-Instruct* (Dubey et al., 2024) is used as the base model in Table 10. The results demonstrate that the min-max curriculum achieves slightly better performance compared to the max and linear curricula.

### C.4 Performance on Base-Answerable and Base-Unanswerable Questions

Table 11 reports the complete results for both the distractor and ideal retrieval settings, evaluated separately on base-answerable and base-unanswerable questions. Notably, in the ideal retrieval setting, answer F1 scores are lower than those observed in the distractor setting. Upon closer analysis, we attribute this trend to limitations of the F1 metric, which can assign partial credit to answers that are incorrect. When we replaced F1 with the exact match (EM) metric, we observed smaller performance gains on the base-unanswerable subsets and more comparable results across the two retrieval settings.

### C.5 Model Performance Grouped by Number of Hops

Table 13 includes the results from Appendix C.1 when grouping by the number of hops required by each question. The results show that as the number of hops increases per question, model performance consistently decreases. However, the models trained with the min-max curriculum still achieve the highest F1 scores.

Setting	Curriculum	HotpotQA			MuSiQue		
		Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1
Distractor	Baseline	60.65	36.47	45.55	25.88	25.35	25.61
	FT-Baseline (Max)	68.52	71.55	70.00	46.06	64.66	53.80
	Linear	72.65	80.53	76.39	47.93	68.45	56.38
	Min-Max	<b>74.97</b>	<b>81.25</b>	<b>77.98</b>	<b>55.13</b>	<b>69.27</b>	<b>61.40</b>
Ideal Ret.	Baseline	67.90	63.26	65.50	41.16	58.16	48.21
	FT-Baseline (Max)	74.79	77.38	76.06	59.04	77.99	67.21
	Linear	77.94	86.45	81.97	64.84	85.23	73.65
	Min-Max	<b>79.74</b>	<b>87.38</b>	<b>83.38</b>	<b>69.79</b>	<b>86.81</b>	<b>77.37</b>

Table 11: Model performance under the distractor and ideal retrieval settings for different curriculum construction settings given larger training set sizes. We use between 20,000-40,000 training samples for all runs as outlined in Appendix C.1.

## D Example Prompts and Outputs

The system prompt and user instructions we use to instruct our models are included in Figures 4 and 5.

**System Prompt**

Respond in the following format:  
 <reasoning>  
 ...  
 </reasoning>  
 <answer>  
 Final answer: final answer  
 Supporting passages: title1, title2, ...  
 </answer>

Figure 4: System prompt used for all experiments.

**User Instructions**

Answer the question using only the provided passages. Verify your answer directly against the text, and cite only the passages you used in your final answer.

Figure 5: User instructions used for all experiments.

Eval. Setting	Curriculum	HotpotQA			MuSiQue			2Wiki		
		Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1
Distractor	Baseline	61.31 / 9.21	44.89 / 35.65	51.83 / 14.64	40.96 / 4.63	27.91 / 19.25	33.20 / 7.46	61.64 / 6.75	42.71 / 33.80	50.45 / 11.25
	FT-Baseline (Max)	82.63 / 18.07	76.01 / 68.70	79.18 / 28.61	63.87 / 13.51	55.54 / 47.51	59.41 / 21.04	84.42 / 17.37	76.53 / 75.50	80.28 / 28.25
	Linear	86.26 / <b>20.48</b>	80.84 / 73.02	83.46 / 31.99	70.18 / 18.82	62.70 / 53.70	66.23 / 27.87	85.00 / <b>23.14</b>	<b>83.77 / 80.32</b>	<b>84.38 / 35.93</b>
	Min-Max	<b>86.27 / 20.40</b>	<b>83.21 / 76.46</b>	<b>84.71 / 32.20</b>	<b>71.73 / 22.89</b>	<b>67.42 / 58.47</b>	<b>69.51 / 32.90</b>	<b>87.83 / 20.41</b>	77.66 / 74.19	82.43 / 32.02
	Base-Answerable	82.63 / 17.09	75.08 / 66.20	78.67 / 27.16	65.62 / 16.17	57.97 / 47.65	61.56 / 24.15	86.26 / 17.42	77.29 / 75.14	81.53 / 28.28
	Base-Unanswerable	82.94 / 16.71	72.74 / 65.35	77.51 / 26.62	63.59 / 14.49	55.44 / 45.81	59.23 / 22.01	83.92 / 16.16	71.42 / 73.19	77.17 / 26.48
Ideal Retrieval	Baseline	75.17 / 5.15	66.03 / 64.70	70.30 / 9.55	60.77 / 4.66	52.22 / 52.02	56.17 / 8.55	70.29 / 6.23	54.49 / 54.49	61.39 / 11.19
	FT-Baseline (Max)	89.74 / 8.61	88.51 / 86.67	89.12 / 15.67	77.67 / 12.14	69.11 / 70.15	73.14 / 20.69	88.64 / 13.56	79.84 / 84.54	84.01 / 23.36
	Linear	90.27 / <b>10.66</b>	91.13 / 89.32	90.70 / <b>19.04</b>	81.70 / 19.65	74.24 / 75.10	77.79 / 31.14	90.50 / <b>19.70</b>	<b>88.34 / 91.37</b>	<b>89.41 / 32.41</b>
	Min-Max	<b>92.06 / 10.30</b>	<b>95.12 / 93.68</b>	<b>93.56 / 18.56</b>	<b>86.46 / 29.55</b>	<b>81.60 / 81.64</b>	<b>83.96 / 43.39</b>	<b>91.01 / 17.17</b>	80.95 / 83.03	85.69 / 28.46
	Base-Answerable	89.74 / 8.87	85.27 / 83.83	87.45 / 16.04	79.11 / 14.17	67.38 / 67.07	72.77 / 23.39	89.68 / 13.76	80.61 / 85.46	84.90 / 23.70
	Base-Unanswerable	<b>88.89 / 7.77</b>	<b>84.98 / 83.93</b>	<b>86.89 / 14.22</b>	<b>77.66 / 12.12</b>	<b>68.00 / 70.96</b>	<b>72.51 / 20.71</b>	<b>88.95 / 11.88</b>	<b>75.07 / 81.82</b>	<b>81.42 / 20.75</b>

Table 12: Model performance under the distractor and ideal retrieval evaluation settings evaluated on **base-answerable** and **base-unanswerable** evaluation sets. We use up to 5,000 training samples for all runs as outlined in Section 4.2

Curriculum	MuSiQue 2-hop			MuSiQue 3-hop			MuSiQue 4-hop		
	Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1
Baseline	28.94	29.46	29.19	23.09	22.62	22.85	21.65	17.80	19.53
FT-Baseline (Max)	48.95	70.47	57.77	44.94	63.84	52.75	39.22	48.23	43.26
Linear	52.04	74.38	61.23	45.53	67.08	54.24	39.74	<b>52.74</b>	45.33
Min-Max	<b>57.03</b>	<b>76.27</b>	<b>65.26</b>	<b>54.11</b>	<b>67.51</b>	<b>60.08</b>	<b>51.16</b>	50.94	<b>51.05</b>

Table 13: Model performance on MuSiQue in the *distractor setting* grouped by the number of hops in each question. We use between 20,000-40,000 training samples for all the runs in this table.