ARMOR: Adaptive Curriculum Meta-Learning for Noise-Robust RAG Reasoning

Yan Wang[1,2], **Yuxin Zhang**[1,2] **Shenyu Zhang**[1,2], **Yongrui Chen**[1,2], **Sheng Bi**[3] and **Guilin Qi**[1,2]
[1]School of Computer Science and Engineering, Southeast University, China
[2]Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3]School of Law, Southeast University, China
{yanwangstu, zzyx_cs, gqi}@seu.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) systems have demonstrated remarkable effectiveness in mitigating hallucinations by incorporating external knowledge. However, the retrieval process inevitably introduces noise, posing significant challenges to RAG robustness. Fundamentally, noise robustness is a composite competency encompassing noise identification, retrieval filtering, and robust reasoning. Unfortunately, existing methods often fail to capture this holistic nature, typically restricted to single-dimensional enhancements or hampered by noise-induced optimization interference. To address these limitations, we decompose RAG robustness into three sub-tasks: knowledge boundary perception, retrieval verification, and robust reasoning. Accordingly, we propose ARMOR, an Adaptive Curriculum Meta-Learning for Noise-Robust RAG Reasoning. Inspired by the teaching process, ARMOR introduces an adaptive task scheduler as a *Teacher* to curate training plans according to the absolute learning progress of RAG generators (*Student*). Distinct from GRPO, ARMOR utilizes inter-group advantage estimation during learning to effectively mitigate mode collapse caused by limited output spaces. Experiments on three complex reasoning datasets (*2WikiMultiHopQA*, *MusiQue*, and *HotPotQA*) indicate that ARMOR achieves synergistic improvements in retrieval noise robustness, significantly outperforming baselines in output quality.

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a critical paradigm for mitigating hallucinations in Large Language Models (LLMs) by incorporating external knowledge [Lewis *et al.*, 2020; Gao *et al.*, 2023]. However, the retrieval process is inherently imperfect and inevitably introduces irrelevant or misleading *Retrieval Noise*, which poses a severe challenge to the robustness of RAG systems [Cuconasu *et al.*, 2024; Wu *et al.*, 2025]. Fundamentally, noise robustness is not a standalone trait but a composite capability. It requires generative systems to simultaneously perform across three distinct dimensions: identifying knowledge boundaries, filtering noise from retrieved context, and executing correct reasoning [Thakur *et al.*, 2024].

Unfortunately, current methods often focus on specific facets of noise robustness [Chen *et al.*, 2024]. Some studies employ supervised fine-tuning to train generators to detect noise and execute robust reasoning within noisy contexts [Zhou *et al.*, 2025; Wei *et al.*, 2025]. Conversely, others leverage generative systems' intrinsic knowledge boundary perception to reduce unnecessary retrieval [Su *et al.*, 2024; Wang *et al.*, 2023]. However, these methods generally restrict optimization to isolated capabilities, lacking the synergy required for comprehensive robustness. Reinforcement Learning (RL) offers a paradigm to address this by enabling the co-evolution of these capabilities via end-to-end policy optimization [Shao *et al.*, 2024; Feng *et al.*, 2025].

Nevertheless, applying standard RL to enhance retrieval noise robustness poses two critical challenges. First, the reward signals derived from the multi-dimensional optimization objectives of noise robustness may result in mutual interference or cancellation, thereby preventing the generative system from converging to an optimal policy for any specific capability [Li *et al.*, 2025]. Second, these capabilities exhibit varying degrees of complexity and follow an underlying sequential logic. Just as *a student must learn addition and subtraction before multiplication and division*, the optimization of noise robustness necessitates a sequential learning trajectory. However, standard RL algorithms typically disregard this inherent progression, attempting to optimize all noise robustness capabilities simultaneously [Huang *et al.*, 2025]. This simultaneous learning strategy often induces an overwhelming optimization burden, restricting the generative system from effectively learning these capabilities.

To address these challenges, we propose ARMOR, an Adaptive Curriculum Meta-Learning for Noise-Robust RAG Reasoning. ARMOR leverages curriculum-based Reinforcement Learning to enhance the comprehensive noise robustness of generative systems. Accordingly, we decouple the composite objective of noise robustness into three sub-tasks: (1) knowledge boundary perception, (2) retrieval verification, and (3) robust reasoning. Specifically, knowledge boundary perception determines whether the model's intrinsic knowl-

edge suffices or external retrieval is required. Retrieval verification distinguishes relevant evidence from irrelevant noise within the retrieved context. Robust reasoning integrates valid information to generate correct answers even in noisy contexts. Systematically acquiring these capabilities is essential for establishing a resilient defense against retrieval noise.

We draw inspiration from the human teaching process, where teachers dynamically adjust the curriculum based on the student's learning feedback [Portelas *et al.*, 2019]. ARMOR models the training process as an adaptive curriculum meta-learning framework, and introduces an adaptive task scheduler as a *teacher*. In each training step, the teacher dynamically selects a specific sub-task for the student (the RAG generator), leveraging Absolute Learning Progress (ALP) to quantify the student's real-time proficiency and prioritize the capability with the highest potential for improvement.

By prioritizing the most critical capability at each stage, ARMOR allows the optimizer to focus on a single dimension of robustness, effectively avoiding the mutual interference and cancellation of reward signals. To further stabilize this decoupled training, we introduce inter-group advantage estimation. Distinct from standard group relative policy optimization (GRPO) [Shao *et al.*, 2024], inter-group advantage estimation extends advantage estimation to a batch-level global scope. By combining inter-group advantage estimation with the adaptive curriculum meta-learning framework, ARMOR effectively eliminates task heterogeneity between different inputs and mitigates mode collapse stemming from limited output spaces. Experiments on three complex reasoning datasets (*2WikiMultiHopQA* [Ho *et al.*, 2020], *MusiQue* [Trivedi *et al.*, 2022], and *HotPotQA* [Yang *et al.*, 2018]) indicate that ARMOR achieves synergistic improvements in retrieval noise robustness, significantly outperforming baselines in output quality.

We summarize our contributions as follows:

- We propose ARMOR, utilizing an ALP-based adaptive task scheduler to dynamically adjust the learning curriculum based on student's feedback.
- We introduce inter-group advantage estimation to stabilize training and mitigate mode collapse stemming from limited output spaces.
- Experiments on three noise-injected complex reasoning datasets demonstrate that ARMOR achieves synergistic improvements in retrieval noise robustness.

2 Related Work

Noise Robustness in RAG. Enhancing the robustness of RAG systems against retrieval noise has become a pivotal research direction. A primary defense against noise involves explicit mechanisms to evaluate and filter retrieved documents before or during generation. CRAG [Yan *et al.*, 2024] employs a lightweight retrieval evaluator to assess document quality, triggering corrective actions when the confidence is low. Similarly, TRUSTRAG [Zhou *et al.*, 2025] utilizes clustering and self-assessment to identify and filter out adversarial or misleading information. Rather than filtering inputs, CHAIN-OF-NOTE [Yu *et al.*, 2024] trains models to generate sequential reading notes, explicitly assessing document pertinence before formulating an answer. INSTRUCTRAG [Wei *et al.*, 2025] further advances this by instructing models to synthesize rationales that explain how the answer is derived from noisy contexts. Additionally, RAAT [Fang *et al.*, 2024] employs adaptive adversarial training, exposing the model to various noise types to internalize robust reasoning capabilities. SKR [Wang *et al.*, 2023] utilizes the model's self-knowledge to classify questions as "known" or "unknown" before invoking a retriever. Similarly, DRAGIN [Su *et al.*, 2024] introduces Real-time Information Needs Detection to trigger retrieval only when the model exhibits uncertainty during generation. However, existing methods typically address noise robustness as an isolated capability by focusing solely on filtering, reasoning, or retrieval triggering.

RL for Robust RAG. Recent advancements have increasingly leveraged RL to enhance the robustness of RAG systems against retrieval noise. Some studies focus on optimizing the interaction policy between reasoning and retrieval. For instance, SEARCH-R1 [Jin *et al.*, 2025] and R1-SEARCHER++ [Song *et al.*, 2025] utilize outcome-based RL to train models to autonomously search queries within reasoning steps, enabling dynamic knowledge acquisition. To enhance noise filtering, R3-RAG [Li *et al.*, 2025] employs a relevance-based process reward to verify retrieved documents at each step. Similarly, LETS [Zhang *et al.*, 2025a] integrates knowledge redundancy and matching rewards to penalize irrelevant retrieval and prevent error propagation in multi-hop reasoning. EVINOTE-RAG [Dai *et al.*, 2025] restructures the workflow into a "retrieve-note-answer" paradigm, using an evidence quality reward to filter noise and distill key evidence before generation. Furthermore, RAG-RL [Huang *et al.*, 2025] investigates static curriculum strategies to improve model robustness against distractors. In contrast, ARMOR decomposes robustness into three distinct sub-tasks, utilizing an ALP-based adaptive task scheduler to dynamically adjust the learning curriculum based on the generator's feedback.

3 Preliminaries

Task Definition: We focus on the task of RAG under noisy retrieval conditions. Given a QA pair $(q, a) \in \mathcal{T}_{train}$, the RAG system utilizes a retriever to fetch a set of external documents $\mathcal{D} = \{d_1, d_2, \cdots, d_k\}$ based on the question q from a large corpus. The retrieved set $\mathcal{D}$ inevitably contains noise due to the inherent imperfections of the retriever. We denote the subset of golden documents as $\mathcal{D}^+ \subseteq \mathcal{D}$ and the noisy documents as $\mathcal{D}^- = \mathcal{D} \setminus \mathcal{D}^+$. The robust RAG generator π_θ aims to generate the correct answer a by effectively filtering $\mathcal{D}^-$ and reasoning over $\mathcal{D}^+$ in conjunction with its intrinsic knowledge. We formulate this as maximizing the probability of the correct answer a given the query q and retrieved set $\mathcal{D}$:

$$\theta^* = \arg\max_{\theta} \sum_{(q,a)\in\mathcal{T}_{train}} \log P_\theta(a \mid q, \mathcal{D}). \quad (1)$$

Warm-Up: To equip the RAG generator with the fundamental capabilities of problem decomposition and step-by-step reasoning, we employ SFT on reconstructed multi-hop RAG datasets as a warm-up stage. This phase lays the

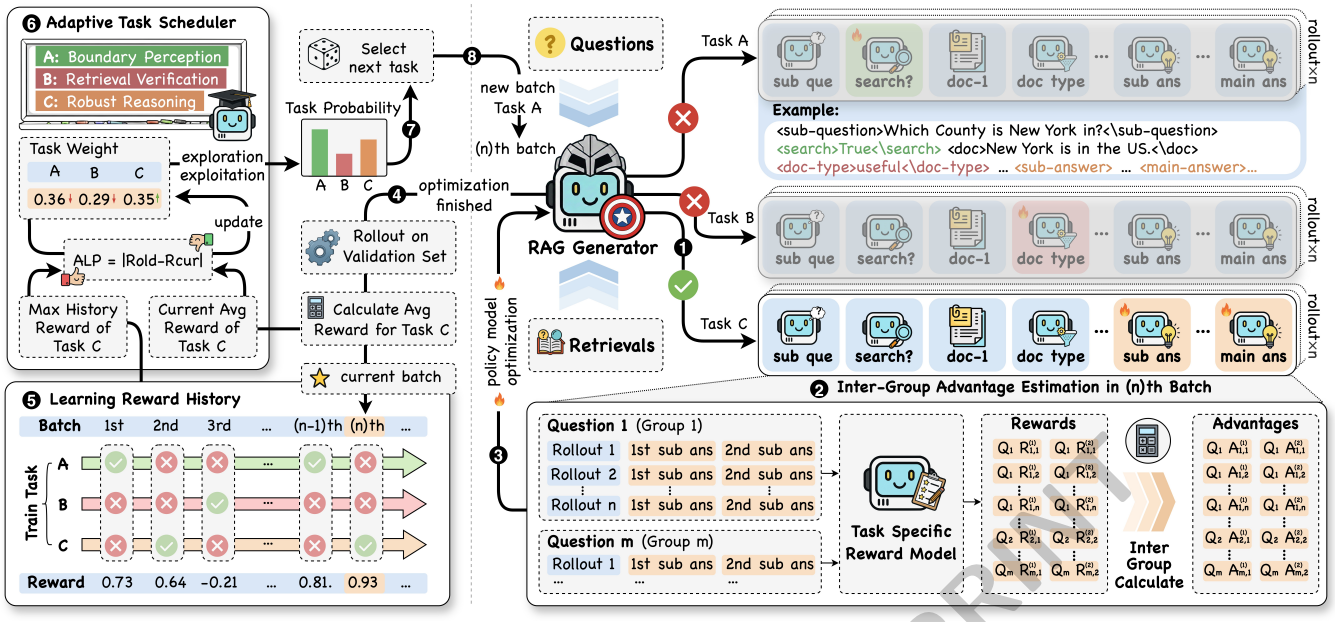

Figure 1: The overall framework of ARMOR. (**Left**) The Adaptive Task Scheduler functions as a *Teacher*, dynamically prioritizing one of the three robustness sub-tasks (Knowledge Boundary Perception, Retrieval Verification, and Robust Reasoning) for each training batch based on the Absolute Learning Progress (ALP). (**Right**) The RAG Generator (*Student*) generates multiple rollouts to learns the selected task. (**Bottom Right**) Inter-Group Advantage Estimation computes reward advantages across the global batch scope rather than within groups, effectively mitigating mode collapse and stabilizing the decoupled training process.

foundation for the subsequent RL phase. For each pair of `<main-question>` and `<main-answer>`, we construct a reasoning chain to serve as the training target. A single reasoning chain consists of structured nodes: `<think>`, `<sub-question>`, `<search>`, `<doc>`, `<doc-type>`, and `<sub-answer>`. Specifically, we train the generator to decompose the `<main-question>` into sub-questions, determine retrieval policies via `<search>True<search>` or `<search>False<search>`, verify retrieved documents using `<doc-type>` tags, and generate corresponding sub-answers based on the reasoning process. This structured supervision ensures that the generator explicitly internalizes the "retrieve-verify-reason" workflow before exploring complex noisy environments. Detailed dataset construction procedures are provided in Appendix A.

4 Methodology

To achieve comprehensive noise robustness, we propose ARMOR, an Adaptive Curriculum Meta-Learning framework. As illustrated in Figure 1, ARMOR consists of two core components: (1) Adaptive Task Scheduler that dynamically curates the learning curriculum based on the learning feedback, and (2) Inter-Group Advantage Estimation mechanism that stabilizes optimization across heterogeneous sub-tasks.

4.1 Adaptive Curriculum Meta-Learning

We decouple the composite objective of noise robustness into three sub-tasks: Knowledge Boundary Perception (τ_{ret}), Retrieval Verification (τ_{fil}), and Robust Reasoning (τ_{ans}). To avoid the mutual interference of reward signals from distinct tasks, we employ an adaptive task scheduler. In each training round n, the scheduler selects a task τ_s from the task set $\mathcal{T} = \{\tau_{\text{ret}}, \tau_{\text{fil}}, \tau_{\text{ans}}\}$ for optimization, prioritizing tasks with high potential for improvement based on learning feedback. We initialize the task weights $w_1(\tau) = 1/|\mathcal{T}|$ for all $\tau \in \mathcal{T}$. In the n-th round, the selection probability $p_n(\tau)$ is computed to balance exploration and exploitation:

$$\gamma = \min\left\{1, \sqrt{\frac{|\mathcal{T}|\ln(|\mathcal{T}|T)}{T}}\right\}, \tag{2}$$

$$p_n(\tau) = (1-\gamma)\frac{w_n(\tau)}{\sum_{\tau'\in\mathcal{T}} w_n(\tau')} + \frac{\gamma}{|\mathcal{T}|}, \tag{3}$$

where T is the predefined training rounds, and γ serves as an exploration coefficient to prevent premature convergence to local optima. The scheduler samples a task $\tau_s \sim p_n(\cdot)$ at the round n. Upon completing the training for the selected task τ_s, we quantify the model's increase in competence using Absolute Learning Progress (ALP) [Portelas *et al.*, 2019]. A higher ALP indicates that the task offers significant potential for performance improvement. The raw ALP is defined as:

$$ALP(\tau) = \mathbb{I}[\tau = \tau_s] \cdot \left|r_n(\tau_s) - r_{\text{old}}(\tau_s)\right|, \tag{4}$$

where $r_n(\tau_s)$ denotes the task reward received in the current training round n (detailed in section 4.3), $r_{\text{old}}(\tau_s)$ represents the maximum historical reward achieved for task τ_s. To correct for the bias introduced by probabilistic sampling, we compute the unbiased reward estimation $\hat{r}_n(\tau)$:

$$\hat{r}_n(\tau) = \frac{ALP(\tau)}{p_n(\tau)}, \tag{5}$$

detailed proofs are provided in Appendix B.2. Finally, the scheduler updates the task weights for the next round, incorporating a forgetting factor $\alpha = 1/T$ to smooth updates and retain historical experience:

$$w_{n+1}(\tau) = w_n(\tau) \cdot \exp\left(\frac{\gamma \hat{r}_n(\tau)}{|\mathcal{T}|}\right) + \frac{e\alpha}{|\mathcal{T}|} \cdot \sum_{\tau' \in \mathcal{T}} w_n(\tau'). \tag{6}$$

4.2 Inter-Group Advantage Estimation

Standard Group Relative Policy Optimization (GRPO) normalizes advantages within a group of rollouts corresponding to a single input. However, intra-group normalization results in unstable learning signals for individual tasks during rollouts. In our decoupled training setting, the constrained output space induces high similarity among rollouts generated from the same input, which reduces intra-group diversity and exacerbates the risk of mode collapse. To overcome this limitation, ARMOR introduces Inter-Group Advantage Estimation, which performs advantage normalization across the entire training batch rather than within isolated inputs.

Given a batch of questions $\mathcal{Q} = \{q_1, \ldots, q_m\}$, we generate l rollouts for each question, denoted as:

$$\mathcal{O}_{\text{train}} = \{o_{i,j} \mid 1 \le i \le m,\ 1 \le j \le l\}, \tag{7}$$

where $o_{i,j}$ denotes the j-th rollout of question q_i. For the selected task τ_s, each rollout $o_{i,j}$ receives a sequence of task-level rewards, denoted as:

$$\mathbf{r}_{i,j} = \left[r_{i,j}^{(1)}, \ldots, r_{i,j}^{(K_{i,j})}\right]^\top, \tag{8}$$

where $r_{i,j}^{(k)}$ represents the reward of k-th task span in the rollout $o_{i,j}$. In addition, each rollout sequence $o_{i,j}$ is associated with a format reward $r_{i,j}^{\text{fmt}}$, which evaluates whether the rollout sequence conforms to the required format. We define the global sets of task and format rewards in current batch as:

$$R_{\text{task}} = \{\mathbf{r}_{i,j} \mid 1 \le i \le m,\ 1 \le j \le l\}, \tag{9}$$

$$R_{\text{fmt}} = \left\{r_{i,j}^{\text{fmt}} \;\middle|\; 1 \le i \le m,\ 1 \le j \le l\right\}. \tag{10}$$

As shown in Figure 1, we normalize rewards using the global statistics of the entire batch to compute the task-level advantage $A_{i,j}^{\text{task}(k)}$ and format-level advantage $A_{i,j}^{\text{fmt}}$:

$$A_{i,j}^{\text{task}(k)} = \frac{r_{i,j}^{(k)} - \text{mean}(R_{\text{task}})}{\text{std}(R_{\text{task}})}, \tag{11}$$

$$A_{i,j}^{\text{fmt}} = \frac{r_{i,j}^{\text{fmt}} - \text{mean}(R_{\text{fmt}})}{\text{std}(R_{\text{fmt}})}, \tag{12}$$

where $\text{mean}(\cdot)$ and $\text{std}(\cdot)$ compute the mean and standard deviation over the entire batch. By combining inter-group advantage estimation with the adaptive curriculum meta-learning framework, ARMOR effectively eliminates task heterogeneity between different inputs and provides a more stable training signal.

The final token-level inter-group advantage $A_{i,j}^{(t)}$ combines dual-level signals:

$$A_{i,j}^{(t)} = \sum_{k=1}^{K_{i,j}} \mathbb{I}[t \in \text{tokens}(i,j,k)] \cdot A_{i,j}^{\text{task}(k)} + \omega A_{i,j}^{\text{fmt}} \tag{13}$$

where t is the token index in rollout $o_{i,j}$, $\text{tokens}(i,j,k)$ denotes the token index of the k-th task span in rollout $o_{i,j}$. $A_{i,j}^{(t)}$ represents the token advantage assigned to token t in rollout $o_{i,j}$. The weighting coefficient ω controls the relative strength of the format signal. Then the clipped policy optimization objective of ARMOR is as follows:

$$\begin{aligned}\mathcal{J}_{\text{ARMOR}}(\theta) = \mathbb{E}\left[q_i \in \mathcal{Q}, \{o_{i,j}\}_{j=1}^{l} \sim \pi_{\theta_{old}}(\cdot \mid qi)\right] \frac{1}{\sum_{j=1}^{l} |o_{i,j}|} \sum_{j=1}^{l} \sum_{t=1}^{|o_{i,j}|} \\ \min\left[\frac{\pi_\theta\left(o_{i,j}^{(t)} \mid q, o_{i,j}^{(<t)}\right)}{\pi_{\theta_{old}}\left(o_{i,j}^{(t)} \mid q, o_{i,j}^{(<t)}\right)} A_{i,j}^{(t)},\right. \\ \left.\text{clip}\left(\frac{\pi_\theta\left(o_{i,j}^{(t)} \mid q, o_{i,j}^{(<t)}\right)}{\pi_{\theta_{old}}\left(o_{i,j}^{(t)} \mid q, o_{i,j}^{(<t)}\right)}, 1-\varepsilon_{low}, 1+\varepsilon_{high}\right) A_{i,j}^{(t)}\right].\end{aligned} \tag{14}$$

After the training for the task τ_s in the n-th batch, the task reward $r_n(\tau_s) = \text{mean}(R_{\text{task}})$ is calculated on the validation dataset, and returned to the adaptive task scheduler to update the task weights.

4.3 Reward Modeling

To precisely guide the generator toward robust behaviors, we design tailored reward functions for each sub-task, addressing their specific optimization challenges.

Knowledge Boundary Perception. The primary objective of this task is to enable the generator to accurately perceive the boundaries of its intrinsic knowledge. For each generated sub-question `<sub-question>`, the RAG generator predicts a `<search>` tag to determine whether to invoke the retriever. The action space is defined as:

$$\mathcal{C}_{retriever} = \{\text{true}, \text{false}\} \tag{15}$$

To align the model's retrieval behavior with the actual necessity of external information, we design a reward function that evaluates whether the model should retrieve from external knowledge based on the correctness of its direct answer. We denote the correctness of the answer generated without retrieval as $v_{\text{ans}} \in \{\text{correct}, \text{error}\}$, and the model's chosen retrieve action as a_{ret}. The reward r_{ret} is defined as follows:

$$r_{\text{ret}} = \begin{cases} 1 & \text{if } v_{\text{ans}} \text{ is error} \wedge a_{\text{ret}} = \text{true}, \\ -1 & \text{if } v_{\text{ans}} \text{ is error} \wedge a_{\text{ret}} = \text{false}, \\ -0.5 & \text{if } v_{\text{ans}} \text{ is correct} \wedge a_{\text{ret}} = \text{true}, \\ 0.5 & \text{if } v_{\text{ans}} \text{ is correct} \wedge a_{\text{ret}} = \text{false}. \end{cases} \tag{16}$$

Retrieval Verification. This task requires the generator to classify each retrieved document into three categories: $\mathcal{C}_{doc} = \{\text{golden}, \text{inconsequential}, \text{irrelevant}\}$ [Zhang *et al.*, 2025b]. Standard coarse-grained binary rewards (correct/incorrect classification) often lead to training instability. To mitigate this, we propose a fine-grained Retrieval Verification Reward based on the probability margin. For a document d, let c^* be the ground-truth label and $p(c)$ be the predicted

probability of class c. We first calculate the confidence margin $m(d)$ between the target class and the most competitive incorrect class:

$$m(d) = p(c^*) - \max_{c \neq c^*, c \in \mathcal{C}_{doc}} p(c). \tag{17}$$

The final retrieval verification task reward is:

$$r_{\text{fil}}(d) = \tanh\big(\beta \cdot m(d)\big), \tag{18}$$

where scaling factor β is used to control the sensitivity of the reward scaling.

Robust Reasoning. The goal of robust reasoning is to generate correct sub-answers `<sub-answer>` and a final answer `<final-answer>` given noisy context. We compute the answer reward r_{ans} by first obtaining a raw correctness score s_{ans} from an LLM evaluator, then linearly mapping it to the range $[-1, 1]$:

$$r_{\text{ans}} = 2 \cdot \frac{s_{\text{ans}} - s_{\min}}{s_{\max} - s_{\min}} - 1, \tag{19}$$

where $[s_{\min}, s_{\max}]$ is the range of correctness score. Details of the LLM-based evaluation are provided in Appendix E.

5 Experiment

In this section, we present empirical evaluations of the effectiveness of ARMOR in defending against retrieval noise across three complex reasoning datasets. Specifically, we aim to validate the following: (1) ARMOR's capability in enhancing the noise robustness of RAG generators; (2) the contribution of core components via ablation studies; (3) the dynamic trends of learning feedback $p_n(\tau)$ during training.

5.1 Experiment Setup

Datasets. As detailed in Appendix A, we reconstruct *2WikiMultiHopQA* and *MuSiQue* datasets. For each dataset, training set consists of two stages: a warm-up stage and a RL stage. The validation set derives from the reconstructed datasets, while the test set consists of two reconstructed subsets and *HotpotQA*, which serves as an out-of-distribution dataset for evaluating model generalization. The detailed dataset statistics are presented in Table 1.

Subset	*2Wiki*	*MuSiQue*	*HotpotQA*
Train$_{\text{Warm-Up}}$	4,467	496	–
Train$_{\text{RL}}$	7,716	857	–
Test	1,352	375	500
Validation	1,352	375	–

Table 1: Experiment dataset statistics.

Evaluation Metrics. We evaluated the model using the following metrics: (1) **Exact Match (EM)** measures the strict accuracy by determining if the predicted answer matches the ground truth verbatim. (2) **F1-Score (F1)** captures the token-level overlap between prediction and reference, offering flexibility for answers with minor phrasing variations. However, these traditional metrics rely on surface-level matching and often fail to capture semantic validity in generative tasks. To address this, we introduce (3) **Correctness (Cor.)**, an LLM-based metric utilizing `GPT-5` to assess the factual accuracy and completeness of answers against references. For clarity, correctness scores are normalized to the range $[0, 1]$. Detailed evaluation prompts and results aligned with manual annotations are presented in the Appendix E.

Baselines. We compare ARMOR against several established baselines: (1) **Adaptive Retrieval**: We compare against SKR [Wang *et al.*, 2023] and DRAGIN [Su *et al.*, 2024], which selectively trigger retrieval based on model's intrinsic knowledge boundary to mitigate noise introduction. (2) **Robust Reasoning**: This category includes CHAIN-OF-NOTE [Yu *et al.*, 2024] and IRCOT [Trivedi *et al.*, 2023]. We employ standard COT without retrieval as a lower bound to benchmark pure reasoning robustness. (3) **RL-Enhanced RAG**: This includes methods like RAG-RL [Huang *et al.*, 2025] and SEARCH-R1 [Jin *et al.*, 2025] that optimize retrieval and reasoning policies via reinforcement learning. In addition, we include NO-RAG (direct inference) and VANILLA-RAG (standard RAG workflow), which serve as performance lower bounds, as they lack specific mechanisms to defend against retrieval noise. We employ two mainstream LLM series with varying parameter scales as the RAG generators: Llama-3.2-1B, Llama-3.2-3B [Team, 2024], Qwen-3-0.6B, Qwen-3-1.7B and Qwen-3-4B [Yang *et al.*, 2025]. For all baselines, we strictly follow the hyperparameter settings and configurations reported in their original studies. Detailed baseline configurations are provided in Appendix C.

Training Details. We train the RAG generator for one epoch using Adam optimizer [Kingma and Ba, 2015] with a learning rate of 5e-5. One epoch includes 857 training rounds, and each round consists of 10 questions. The clipping coefficient ε_{low} and ε_{high} are set to 0.2 and 0.28, separately. The default rollout number l is set to 2. The weighting coefficient ω is set to 1, and the scaling factor β in the retrieval verification task is set to 6. During the rollout process, we employ `e5-base-v2` [Wang *et al.*, 2022] as the retriever, retrieving a maximum of three documents per retrieval triggering. For generation, we adopt a sampling temperature of 0.85, along with top-$p = 0.95$ and top-$k = 50$. All reported results are averaged over three independent runs to ensure statistical reliability.

5.2 Main Results

As demonstrated in our preliminary study (Appendix C.2), standard retrieval systems inevitably retrieve irrelevant or inconsequential documents, introducing *natural noise* into the generation process. Table 2 presents the comparative performance of ARMOR and baselines under this realistic setting. ARMOR achieves SOTA performance across all datasets and metrics, demonstrating remarkable consistency on both Qwen and Llama backbones. Specifically, ARMOR surpasses the strongest baseline (IRCOT) by a substantial margin, achieving an average Correctness improvement of 38.8% on Qwen-3-1.7B and 19.9% on Llama-3.2-3B. Standard VANILLA-RAG shows limited improvement over direct inference (NO-RAG) and performs poorly on complex datasets. This confirms that without explicit robustness mechanisms, generators

Method	*2WikiMultiHopQA*			*MusiQue*			*HotpotQA*			*Avg.*		
	Cor.	**F1**	**EM**	**Cor.**	**F1**	**EM**	**Cor.**	**F1**	**EM**	**Cor.**	**F1**	**EM**
Qwen-3-1.7B												
NO-RAG	37.72	37.21	32.99	3.73	2.96	0.00	24.20	22.44	15.67	21.88	20.87	16.22
VANILLA-RAG	47.63	44.67	38.46	24.00	27.43	17.33	45.07	45.72	38.33	38.90	39.27	31.37
COT	39.64	17.13	11.39	8.27	6.27	1.33	22.47	18.03	13.60	23.46	13.81	8.77
CON	63.61	13.13	4.73	29.33	12.36	5.33	54.53	22.65	17.07	49.16	16.05	9.04
IRCOT	50.59	48.80	41.86	32.00	29.97	18.67	55.40	51.97	46.26	46.00	43.58	35.60
DRAGIN	39.70	39.46	36.98	6.67	8.87	5.33	26.87	26.52	22.73	24.41	24.95	21.68
SKR	35.50	32.00	26.92	5.33	3.86	0.00	21.13	20.78	13.20	20.65	18.88	13.37
RAG-RL	49.70	46.67	41.39	29.66	28.79	20.33	48.33	47.64	42.67	42.56	41.03	34.80
SEARCH-R1	41.57	39.61	37.28	10.67	10.55	3.73	28.27	28.21	21.93	26.84	26.12	20.98
ARMOR (warm-up stage)	70.41	70.69	68.34	38.67	39.04	29.33	47.53	45.20	38.86	52.20	51.64	45.51
ARMOR	**79.14**	**79.22**	**77.22**	**54.67**	**57.22**	**46.67**	**57.73**	**54.86**	**58.00**	**63.85**	**63.77**	**60.63**
Llama-3.2-3B												
NO-RAG	28.25	30.09	25.15	5.33	2.32	0.00	22.60	17.98	13.40	18.73	16.80	12.85
VANILLA-RAG	59.32	57.58	53.40	16.00	17.48	10.67	50.27	44.52	36.27	41.86	39.86	33.45
COT	36.83	17.22	14.20	8.00	3.82	1.33	35.60	23.72	18.07	26.81	14.92	11.20
CON	51.33	7.80	1.92	14.67	3.71	0.00	47.47	10.63	6.20	37.82	7.38	2.71
IRCOT	61.39	58.76	53.40	34.67	32.43	22.67	60.13	57.77	48.00	52.06	49.65	41.36
DRAGIN	42.60	41.76	39.50	9.33	7.77	4.00	37.47	35.85	29.73	29.80	28.46	24.41
SKR	35.21	36.98	32.99	8.26	6.95	2.67	35.00	32.50	31.33	26.16	25.48	22.33
RAG-RL	64.48	62.58	58.40	29.67	30.21	21.33	56.60	50.47	41.27	50.25	47.75	40.33
SEARCH-R1	42.31	40.76	38.76	6.67	8.50	4.00	26.33	27.22	24.53	25.10	25.49	22.43
ARMOR (warm-up stage)	71.01	71.28	68.93	32.00	36.54	21.33	45.67	47.79	37.60	49.56	51.87	42.62
ARMOR	**77.81**	**78.09**	**75.59**	**42.67**	**45.62**	**29.33**	**66.93**	**66.47**	**58.67**	**62.47**	**63.39**	**54.53**

Table 2: Performance comparing ARMOR with baselines. We report Correctness (Cor.), F1 score, and Exact Match (EM) in percentage (%) using Qwen-3-1.7B and Llama-3.2-3B as backbones. The best results are highlighted in **bold**. Full results are provided in Appendix D.

are easily misled by the inherent noise in natural retrieval. IRCOT remains competitive by interleaving reasoning and retrieval steps. CON struggles to maintain stability, suggesting that their sequential noise-filtering logic is prone to error propagation in open-domain settings. ARMOR's decoupled training objective effectively mitigates this issue. Existing RL-based approaches (RAG-RL, SEARCH-R1) exhibit suboptimal performance, highlighting the difficulty of optimizing RAG generators with sparse rewards under natural noise.

We conducted a controlled experiment by manually injecting noise into the retrieval. We fixed the number of retrieved documents to 5 and varied the noise ratio from 0% to 100%, as shown in Figure 2. Baseline methods exhibit a **precipitous performance decline** as the noise ratio increases. IRCOT and SEARCH-R1 suffer significant losses, indicating they struggle to filter out distracting information effectively when valid evidence is scarce. In contrast, ARMOR demonstrates exceptional stability, maintaining a high correctness score of 69% even at an 80% noise ratio. ARMOR often outperforms the DIRECT ANSWER baseline or remains comparable at 100% noise. This indicates that ARMOR has effectively learned a *rejection mechanism* via **Inter-Group Advantage Estimation**, allowing it to suppress the influence of irrelevant contexts rather than being hallucinated.

5.3 Ablation Study

We compare ARMOR with the following variants: (1) *w/o CRL*: replaces the ALP-based scheduler with random sampling (*-random*), rigid sequential ordering (*-seq*), or outcome-supervised training (*-outcome sup*: removing process rewards); (2) *w/o Inter-Group Adv*: reverts global advantage normalization to standard intra-group normalization; (3) *Base Setting*: removes both adaptive curriculum and inter-group estimation, reverting to standard outcome-supervised RL. As shown in Table 3, ARMOR consistently outperforms all variant settings across all metrics. Specifically, *w/o CRL -seq* suffers a 47.51% drop in average correctness, indicating that rigid task ordering leads to **catastrophic forgetting** among decoupled sub-tasks. Similarly, *w/o CRL -random* underperforms relative to ARMOR, confirming that the ALP-based scheduler is essential for dynamically balancing exploration and exploitation based on learning feedback. Reverting to standard intra-group normalization (*w/o inter-group adv.*) degrades performance by 14.58%. This decline highlights that standard intra-group normalization **yields unstable signals in our decoupled setting**. In contrast, ARMOR's batch-level advantage estimation provides robust supervision, effectively preventing mode collapse and stabilizing optimization.

5.4 Analysis of Dynamic Learning Feedback

We visualize the dynamic trends of the task selection probability $p_n(\tau)$ throughout the training process. Figure 3 illustrates these trends for both Llama-3.2-3B and Qwen-3-1.7B backbones. Observing the training dynamics of Qwen-3-1.7B, the scheduler assigns high probability to the Knowledge Boundary Perception task in the early stages. This in-

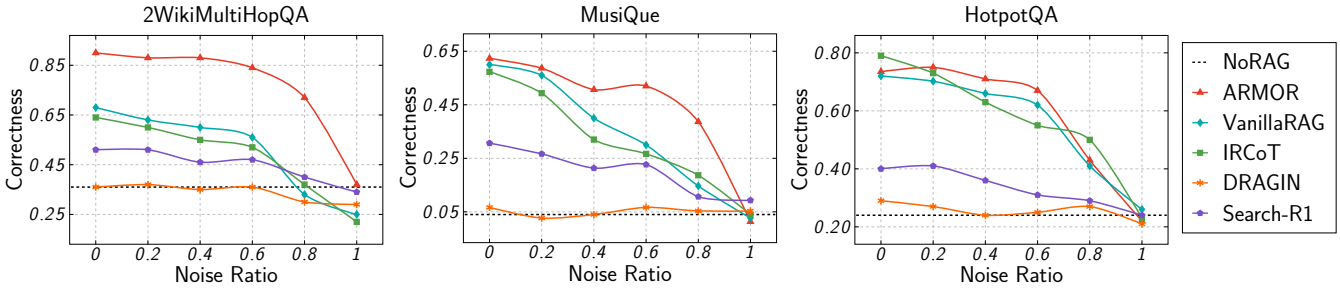

Figure 2: Robustness analysis under controlled noise ratios (from 0% to 100%) with fixed retrieval size of $K = 5$.

Method	*2WikiMultiHopQA*			*MusiQue*			*HotpotQA*			*Avg.*		
	Cor.	**F1**	**EM**	**Cor.**	**F1**	**EM**	**Cor.**	**F1**	**EM**	**Cor.**	**F1**	**EM**
Qwen-3-1.7B												
ARMOR	**79.14**	**79.22**	**77.22**	**54.67**	**57.22**	**46.67**	**57.73**	**54.86**	**58.00**	**63.85**	**63.77**	**60.63**
w/o CRL -random	67.66	67.97	66.48	39.33	40.59	28.67	37.27	35.45	27.33	48.09	48.00	40.83
w/o CRL -seq	52.22	52.69	50.89	23.65	26.50	15.20	24.67	21.76	11.27	33.51	33.65	25.79
w/o CRL -outcome sup.	60.22	60.45	58.59	32.33	32.37	23.00	30.80	27.72	18.73	41.12	40.18	33.44
w/o inter-group adv.	73.55	73.49	71.63	48.00	46.28	40.00	42.07	39.73	31.13	54.54	53.17	47.59
Base Setting	71.63	72.04	70.30	44.33	45.79	36.67	40.47	38.29	30.07	52.14	52.04	45.68

Table 3: Ablation analysis on the effectiveness of the Adaptive Curriculum Learning (CRL) strategy and Inter-Group Advantage Estimation (inter-group adv). The best results are highlighted in **bold**.

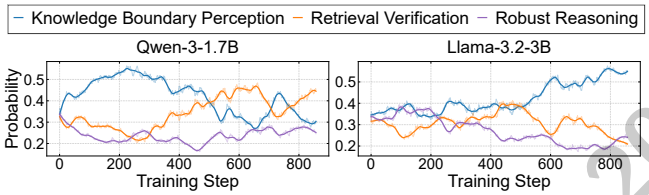

Figure 3: Task selection probabilities $p_n(\tau)$ during the training.

dicates that the generator primarily focuses on distinguishing "*what it knows*" vs. "*what it needs to retrieve.*" The scheduler shifts its focus toward Retrieval Verification around step 300. This transition suggests that once the generator learns "*when to retrieve*," it begins to focus on "*how to filter noise*" from the retrieval. Meanwhile, the attention on Robust Reasoning remains relatively stable, serving as a consistent regularization throughout the process. The trends for Llama-3.2-3B exhibit a different pattern, highlighting the scheduler's adaptability to different generator architectures. Unlike Qwen, Llama shows a continuous and increasing demand for Knowledge Boundary Perception. This suggests that for the Llama backbone, accurately determining the necessity of retrieval is a more challenging task. These findings confirm that ARMOR does not impose a rigid *"one-size-fits-all"* curriculum. Instead, it acts as a dynamic meta-learner, tailoring the training strategy to the learning trajectory of the underlying generator.

5.5 Case Study

We analyze a multi-hop query: "*When was Ole Rynning's father born?*" (answer: "*May 30, 1778*"), detailed case is provided in Appendix D.2. ARMOR successfully decomposes the logic: identifying the father ("*Jens Rynning*") first, then retrieving his birth year. Crucially, ARMOR effectively filters out distractor documents describing the son's birth "*1809*", leading to the correct answer. In contrast, SEARCH-R1 exhibits chaotic reasoning, generating hallucinated contexts and failing to align the answer type. IRCOT suffers from distraction failure, although it retrieves the correct evidence, it lacks explicit noise rejection and is misled by the high semantic overlap of the son's birth date. This case validates that ARMOR's adaptive curriculum meta-learning effectively equips the generator with the ability to distinguish golden evidence from noise in complex reasoning scenarios.

6 Conclusion

In this work, we presented ARMOR, a novel framework designed to fortify RAG systems against the pervasive challenge of retrieval noise. ARMOR decouples the training objective into three distinct sub-tasks: Knowledge Boundary Perception, Retrieval Verification, and Robust Reasoning. To coordinate these tasks, ARMOR introduces an Adaptive Curriculum Meta-Learning strategy, where a *teacher* scheduler dynamically prioritizes learning objectives based on the generator's Absolute Learning Progress. Moreover, ARMOR addresses the optimization instability inherent in decoupled reinforcement learning by introducing Inter-Group Advantage Estimation, which mitigates mode collapse and ensures stable gradient optimization via global batch normalization. Extensive experiments across three complex multi-hop reasoning datasets demonstrate that ARMOR exhibits exceptional resilience, maintaining high correctness even under extreme noise ratios. This confirms that Adaptive Curriculum Scheduling and Inter-Group Advantage Estimation are essential for comprehensive RAG noise-robust reasoning.

Acknowledgments

This work is partially supported by National Nature Science Foundation of China under No. 62476058. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

Contribution Statement

Yan Wang and Yuxin Zhang conceived the study, designed the experiments, and wrote the manuscript. Yan Wang collected the data and performed the experiments. Shenyu Zhang, Yongrui Chen, and Sheng Bi reviewed and edited the final manuscript. Yan Wang and Yuxin Zhang contributed equally to this work. Guilin Qi is the corresponding author.

References

[Chen *et al.*, 2024] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press, 2024.

[Cuconasu *et al.*, 2024] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 719–729. ACM, 2024.

[Dai *et al.*, 2025] Yuqin Dai, Guoqing Wang, Yuan Wang, et al. Evinote-rag: Enhancing RAG models via answer-supportive evidence notes. *CoRR*, abs/2509.00877, 2025.

[Fang *et al.*, 2024] Feiteng Fang, Yuelin Bai, Shiwen Ni, et al. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10028–10039. Association for Computational Linguistics, 2024.

[Feng *et al.*, 2025] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. *CoRR*, abs/2505.10978, 2025.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, et al. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023.

[Ho *et al.*, 2020] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics, 2020.

[Huang *et al.*, 2025] Jerry Huang, Siddarth Madala, Risham Sidhu, et al. RAG-RL: advancing retrieval-augmented generation via RL and curriculum learning. *CoRR*, abs/2503.12759, 2025.

[Jin *et al.*, 2025] Bowen Jin, Hansi Zeng, Zhenrui Yue, et al. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516, 2025.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Li *et al.*, 2025] Yuan Li, Qi Luo, Xiaonan Li, Bufan Li, et al. R3-RAG: learning step-by-step reasoning and retrieval for llms via reinforcement learning. *CoRR*, abs/2505.23794, 2025.

[Portelas *et al.*, 2019] Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, pages 835–853. PMLR, 2019.

[Shao *et al.*, 2024] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

[Song *et al.*, 2025] Huatong Song, Jinhao Jiang, Wenqing Tian, et al. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning. *CoRR*, abs/2505.17005, 2025.

[Su *et al.*, 2024] Weihang Su, Yichen Tang, Qingyao Ai, et al. DRAGIN: dynamic retrieval augmented generation based on the real-time information needs of large language

models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12991–13013. Association for Computational Linguistics, 2024.

[Team, 2024] Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

[Thakur *et al.*, 2024] Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, et al. "knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 12508–12526. Association for Computational Linguistics, 2024.

[Trivedi *et al.*, 2022] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554, 2022.

[Trivedi *et al.*, 2023] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10014–10037. Association for Computational Linguistics, 2023.

[Wang *et al.*, 2022] Liang Wang, Nan Yang, Xiaolong Huang, et al. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022.

[Wang *et al.*, 2023] Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10303–10315. Association for Computational Linguistics, 2023.

[Wei *et al.*, 2025] Zhepei Wei, Wei-Lin Chen, and Yu Meng. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[Wu *et al.*, 2025] Jinyang Wu, Shuai Zhang, Feihu Che, et al. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of RAG noise in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5019–5039. Association for Computational Linguistics, 2025.

[Yan *et al.*, 2024] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *CoRR*, abs/2401.15884, 2024.

[Yang *et al.*, 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, et al. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics, 2018.

[Yang *et al.*, 2025] An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

[Yu *et al.*, 2024] Wenhao Yu, Hongming Zhang, Xiaoman Pan, et al. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14672–14685. Association for Computational Linguistics, 2024.

[Zhang *et al.*, 2025a] Qi Zhang, Shouqing Yang, Lirong Gao, et al. Lets: Learning to think-and-search via process-and-outcome reward hybridization. *CoRR*, abs/2505.17447, 2025.

[Zhang *et al.*, 2025b] Yuxin Zhang, Yan Wang, Yongrui Chen, et al. Magic mushroom: A customizable benchmark for fine-grained analysis of retrieval noise erosion in RAG systems. *CoRR*, abs/2506.03901, 2025.

[Zhou *et al.*, 2025] Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. Trustrag: Enhancing robustness and trustworthiness in RAG. *CoRR*, abs/2501.00879, 2025.