Improving Context Fidelity via Native Retrieval-Augmented Reasoning

Anonymous Authors[1]

Abstract

Large language models (LLMs) often struggle with context fidelity, producing inconsistent answers when responding to questions based on provided information. Existing approaches either rely on expensive supervised fine-tuning to generate evidence post-answer or train models to perform web searches without necessarily improving utilization of the given context. We propose **CARE**, a novel native retrieval-augmented reasoning framework that teaches LLMs to explicitly integrate in-context evidence within their reasoning process with the model's own retrieval capabilities. Our method requires minimal labeled evidence data while significantly enhancing both retrieval accuracy and answer generation performance through strategically retrieved in-context tokens in the reasoning chain. Extensive experiments on multiple real-world and counterfactual QA benchmarks demonstrate that our approach substantially outperforms supervised fine-tuning, traditional retrieval-augmented generation methods, and external retrieval solutions. This work represents a fundamental advancement in making LLMs more accurate, reliable, and efficient for knowledge-intensive tasks.

1. Introduction

Large language models (LLMs) have demonstrated impressive performance in a wide range of natural language tasks (Minaee et al., 2024; Liu et al., 2025a), yet continue to struggle with a fundamental challenge: maintaining fidelity to the context provided when answering questions (Talukdar & Biswas, 2024). This *context hallucination problem* (Chang et al., 2024; Hu et al., 2024; Liu et al., 2025b) is particularly pronounced in knowledge-intensive tasks where precise information retrieval and accurate reasoning are paramount. When LLMs generate answers that contradict or fabricate information relative to the input context, user trust declines, and the practical utility of these systems decreases considerably.

Current approaches to address this challenge fall into two categories with significant limitations. First, retrieval-augmented generation (RAG) methods (Variengien & Winsor, 2023; Wang et al., 2024) improve explainability but require expensive labeled datasets with ground-truth evidence spans, limiting scalability. Second, external retrieval mechanisms (Hsu et al., 2024; Nguyen et al., 2024) access specialized information but underutilize the rich context already provided by users, which often contains the most relevant information for their queries.

In this paper, we introduce **native retrieval-augmented reasoning**, a fundamentally different approach where LLMs dynamically identify and incorporate relevant evidence from input context directly within their reasoning chain, rather than treating retrieval and reasoning as separate processes. This leverages LLMs' inherent language understanding for in-context retrieval without additional indexing or embedding systems while enhancing reasoning through explicit evidence integration. Based on this approach, we introduce the **Context-Aware Retrieval-Enhanced reasoning (CARE)** framework, which requires minimal labeled evidence data and employs two-phase training: (1) supervised fine-tuning (SFT) establishes evidence integration patterns, (2) reinforcement learning (RL) refines self-retrieval through retrieval-aware rewards. Crucially, a curriculum learning schedule enables progressive adaptation from simple to complex reasoning tasks, extending beyond the initial training distribution without additional labeled data.

Our main contributions are as follows.

- We introduce **native retrieval-augmented reasoning**, a novel paradigm that organically combines in-context retrieval with structured reasoning to improve context fidelity and reduce hallucinations.
- We present a curated dataset for training models to perform evidence-integrated reasoning, which we will open source to facilitate further research in this area.
- We propose **CARE**, a comprehensive implementation

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Under review by the Workshop on Long-Context Foundation Models (LCFM) at ICML 2025. Do not distribute.

that combines native retrieval-augmented reasoning with curriculum learning to handle diverse question-answering scenarios without additional labeled data.

- Through extensive experiments across multiple real-world and counterfactual QA benchmarks, we demonstrate that our approach substantially outperforms vanilla SFT, traditional RAG methods, and comparable models lacking in-context retrieval mechanisms in both evidence retrieval and answer accuracy.

2. The CARE Method

2.1. Overview

We present CARE, a reasoning framework that enables LLMs to autonomously perform native retrieval from the input context without external modules, integrating retrieved evidence directly into reasoning instead of outputting them independently. By performing native retrieval, CARE better leverages LLMs' powerful language understanding capabilities while reducing expensive tool calling dependencies. Native retrieval integration improves both context loyalty and reasoning quality through curated evidence.

To minimize reliance on expensive supporting fact labels, CARE employs two-phase training: supervised fine-tuning (SFT) followed by reinforcement learning (RL).

2.2. The Supervised Fine-Tuning Phase

The SFT phase establishes evidence integration by injecting retrieval tokens within structured reasoning steps. Using existing QA datasets with labeled supporting facts, this phase addresses the RL training "cold-start" problem while familiarizing the model with the target output format, native retrieval process, and chain-of-thought reasoning with retrieved evidence.

Our data generation pipeline operates on $\mathcal{D}_{\text{original}} = \{(Q_i, C_i, A_i, S_i)\}_{i=1}^{N_{\text{original}}}$ containing queries, contexts, answers and supporting facts through three sequential stages: reasoning step generation, evidence integration, and retrieval token insertion (Figure 1, top).

Reasoning Step Generation. A reasoning model M_R generates an initial reasoning chain $R_{i,A} = M_R(C_i, Q_i)$. We retain only responses with correct answers and extract reasoning chains N_i from within the ⟨ THINK ⟩ ⟨ /THINK ⟩ tokens.

Evidence Integration. To ground reasoning in context rather than internal knowledge, a non-reasoning model M_I integrates ground-truth supporting facts: $R_{i,I} = M_I(Q_i, N_i, S_i)$. We keep instances where all supporting facts appear in $R_{i,I}$, yielding evidence-grounded chains E_i.

Retrieval Marking. We insert ⟨ RETRIEVAL ⟩ ⟨ /RETRIEVAL ⟩ tokens around evidence segments in E_i to create structured responses E_i^*, which serves as the ground-truth output for the SFT dataset.

The final dataset $\mathcal{D}_{\text{SFT}} = \{(Q_i, C_i, A_i, E_i^*)\}_{i=1}^{N_{\text{SFT}}}$ provides context-grounded reasoning chains with explicit evidence marking for subsequent RL training.

2.3. Reinforcement Learning Phase.

We refine the self-retrieval mechanism from SFT using Group Relative Policy Optimization (GRPO) with curriculum learning to transition from basic to advanced reasoning tasks. A detailed training algorithm is provided in Algorithm 1.

The GRPO algorithm. GRPO evaluates multiple sampled outputs at the group level. Given query q and outputs $\{o_1, \ldots, o_G\}$ sampled from $\pi_{\theta_{old}}$, the objective is:

$$
\begin{aligned}
J_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}} \Bigg[\Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \\
\min \Big[r_{i,t} \hat{A}_{i,t}, \text{clip}\left(r_{i,t}, 1-\epsilon, 1+\epsilon\right) \hat{A}_{i,t} \Big] \\
- \beta D_{\text{KL}}\left(\pi_\theta \parallel \pi_{ref}\right) \Bigg] \Bigg]
\end{aligned} \tag{1}
$$

where $r_{i,t} = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}$ is the importance ratio, clipped to $[1-\epsilon, 1+\epsilon]$. The KL divergence term prevents excessive divergence from the reference policy.

Reward Design. We design three reward components to encourage context-grounded reasoning:

1. **Retrieval Reward** (R_{ret}): Rewards correct use of ⟨ RETRIEVAL ⟩ ⟨ /RETRIEVAL ⟩ pairs when enclosed text exists in the context, enabling dynamic context integration without ground-truth retrieval data.
2. **Format Reward** (R_{fmt}): Ensures structural consistency with both ⟨ THINK ⟩ ⟨ /THINK ⟩ and ⟨ RETRIEVAL ⟩ ⟨ /RETRIEVAL ⟩ pairs.
3. **Accuracy Reward** (R_{acc}): Measures correctness through the token F1 score between the generated and ground-truth answers.

The total reward combines these components: $R_{total} = \lambda_1 R_{acc} + \lambda_2 R_{fmt} + \lambda_3 R_{ret}$, where coefficients $\lambda_1, \lambda_2, \lambda_3$ balance factual precision, structural consistency, and context fidelity.

Curriculum Learning Strategy. QA datasets exhibit significant variation in context and answer lengths. To gradually adapt our model to diverse dataset characteristics other

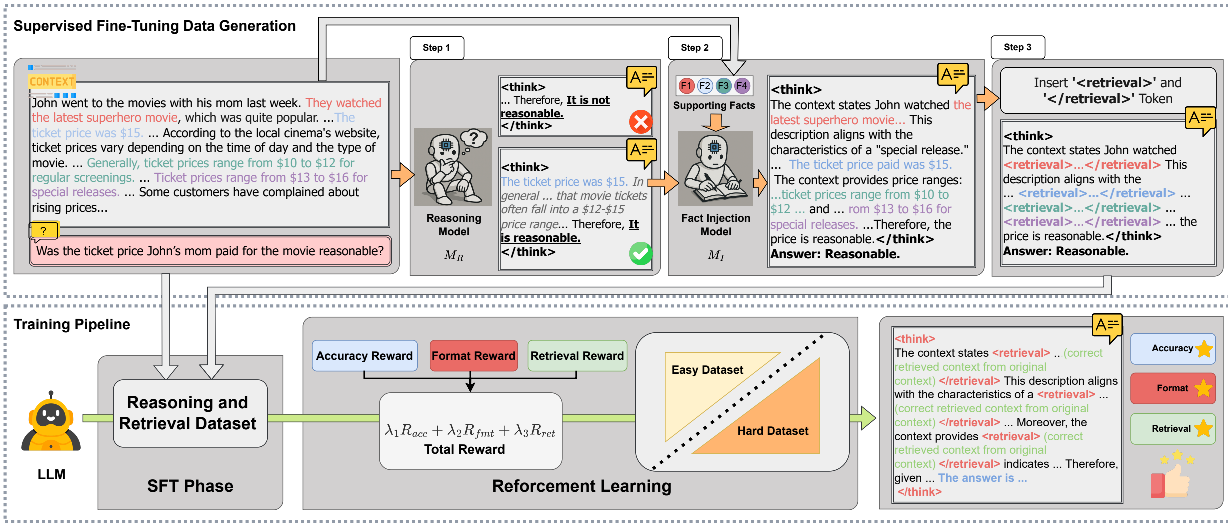

Figure 1. A schematic illustration of the training data and training process. The upper part depicts the SFT data generation pipeline including fact injection and special tokens insertion within the reasoning content, while the lower part illustrates the SFT training process and the reinforcement learning (RL) training with multiple rewards.

Model	Method	MFQA	HotpotQA	2WikiMQA	MuSiQue	Average
LLaMA-3.1 8B	Original	45.57	54.64	45.87	32.08	44.54
	ReSearch	/	/	/	/	/
	R1-Searcher	28.44	53.71	67.10	41.41	47.67
	CRAG	/	/	/	/	/
	CARE	**49.94**	**63.09**	**75.29**	**51.00**	**59.83**
Qwen2.5 7B	Original	46.94	58.47	46.96	30.78	45.79
	ReSearch	32.45	54.24	55.78	**47.61**	47.52
	R1-Searcher	28.36	55.43	65.79	47.09	49.17
	CRAG	47.90	43.97	33.00	28.44	38.33
	CARE	**48.11**	**63.45**	**70.11**	45.57	**56.81**
Qwen2.5 14B	Original	47.58	61.94	59.05	37.99	51.64
	ReSearch	/	/	/	/	/
	R1-Searcher	/	/	/	/	/
	CRAG	**50.89**	44.74	34.68	28.17	39.62
	CARE	48.81	**67.75**	**78.68**	**51.27**	**61.63**

Table 1. Evaluation on the real-world long-sequence QA datasets. The results are grouped by the base LLM used. The best and second-best results for each base model and dataset are labeled in **bold** and underline, respectively. Slash (/) indicates that the method does not have an official checkpoint or support for this model.

Settings	SFT	RL	Ret.	Cur.	MFQA	HotpotQA	2WikiMQA	MuSiQue	CofCA	Average
Baseline	✗	✗	✗	✗	46.64	58.47	46.96	30.78	58.38	48.25
SFT Only	✓	✗	✗	✗	42.24	47.08	61.51	33.82	59.21	48.77
No Ret.	✓	✓	✗	✗	37.66	62.59	70.57	43.85	57.26	54.39
No Cur.	✓	✓	✓	✗	38.33	**64.10**	**70.69**	**47.49**	60.60	56.24
CARE	✓	✓	✓	✓	**48.11**	63.45	70.11	45.57	**64.56**	**58.36**

Table 2. Ablation studies on the QA tasks based on Qwen2.5 7B. The best and second-best results for each base model and dataset are labeled in **bold** and underline, respectively. "Ret." stands for retrieval reward, and "Cur." stands for curriculum learning in Algorithm 1.

than the one used for SFT, we implement a curriculum learning strategy transitioning from short-context / short-answer QA to long-context / multihop long-answer QA. This structured progression mitigates catastrophic forgetting while enhancing retrieval capabilities across multiple task complexity.

Model	Method	CofCA
LLaMA-3.1 8B	Original	<u>48.14</u>
	R1-Searcher	45.25
	CARE	**61.83**
Qwen2.5 7B	Original	<u>58.38</u>
	ReSearch	47.32
	R1-Searcher	43.61
	CRAG	56.01
	CARE	**64.56**
Qwen2.5 14B	Original	<u>64.40</u>
	CRAG	51.99
	CARE	**67.75**

Table 3. Evaluation on the counterfactual QA task CofCA. The results are grouped by the base LLM used. The best and second-best results for each base model and dataset are labeled in **bold** and <u>underline</u>, respectively.

We train with two QA datasets: $\mathcal{D}_{\text{easy}} = \{(Q_i, C_i, A_i)\}_{i=1}^{N_{\text{easy}}}$ and $\mathcal{D}_{\text{hard}} = \{(Q_i, C_i, A_i)\}_{i=1}^{N_{\text{hard}}}$, where $\mathcal{D}_{\text{hard}}$ contains longer contexts, longer answers, and requires more complex reasoning than $\mathcal{D}_{\text{easy}}$. Training begins exclusively with $\mathcal{D}_{\text{easy}}$, then gradually incorporates instances from $\mathcal{D}_{\text{hard}}$.

At each training step t, we sample instances using a Bernoulli trial with a time-varying probability. The mixing ratio α_t decreases linearly according to $\alpha_t = \max(0, 1 - \eta \cdot \frac{t}{T})$, where η is a scaling factor that controls the speed of transition. The sampling probabilities are $p_{\text{easy}} = \alpha_t$ and $p_{\text{hard}} = 1 - \alpha_t$, ensuring the model maintains short-context retrieval capabilities while learning to aggregate evidence across multiple paragraphs.

3. Experiments

We evaluate our proposed CARE method through comprehensive experiments across multiple LLM families and sizes in two distinct QA categories: real-world long-context QA and counterfactual multi-hop QA. Detailed settings are presented in Appendix C.

3.1. Long-Sequence QA Performance

Table 1 shows that CARE consistently outperforms baselines in all model sizes. With LLaMA-3.1 8B, we achieve +15.29% average F1 improvement, with strongest gains on multi-hop tasks (2WikiMQA +29.42%, MuSiQue +18.92%). Qwen2.5 models show similar patterns. When not achieving top performance, CARE remains competitive with the best baselines. These results demonstrate that CARE significantly enhances performance by effectively integrating in-context evidence during reasoning, particularly for complex multi-hop questions. Appendix D shows that CARE also achieves significantly higher evidence retrieval accuracy on HotpotQA compared to the baselines.

3.2. Counterfactual QA Performance

In Table 3, we report the results on the CofCA counterfactual QA task. CARE consistently delivers the strongest performance, with significant gains on LLaMA-3.1 8B (+13.69%). In particular, traditional online search methods underperform compared to original models on this task, suggesting that external retrieval can be counterproductive when context contradicts parametric knowledge. CARE demonstrates superior context fidelity by explicitly integrating natively extracted in-context evidence in the reasoning process, and can make even greater gains compared to the baselines when encountering unseen information in the context.

3.3. Ablation Studies

Table 2 presents the ablation results in Qwen2.5 7B in three settings: (1) **SFT only** (without the RL training phase), (2) **No retrieval reward** (GRPO with DeepSeek-R1-like reasoning reward without retrieval reward), and (3) **No curriculum learning** (RL in $\mathcal{D}_{\text{easy}}$ only).

SFT alone provides marginal gains, while RL training substantially improves performance, confirming reinforcement learning's importance for QA reasoning. Both native in-context reasoning methods ("No Cur." and CARE) consistently outperform vanilla R1-like GRPO ("No Ret."), demonstrating that retrieval-augmented reasoning improves performance by grounding reasoning in contextual evidence. While "No Cur." excels on multihop datasets, curriculum learning achieves better balance across diverse QA types, particularly improving long-form answering (MFQA) and counterfactual scenarios (CofCA). This shows that curriculum learning successfully adapts the model to various types of question while maintaining strong complex reasoning performance, all without additional labeled data beyond the initial SFT.

4. Conclusion

We introduce CARE, a native retrieval-augmented reasoning framework that improves context fidelity in LLMs by teaching models to dynamically identify and integrate evidence within their reasoning process. This approach improves how LLMs interact with context while requiring minimal labeled evidence. Experiments on multiple general and counterfactual QA benchmarks demonstrated that CARE consistently outperforms existing approaches, including the vanilla SFT method and traditional RAG methods on both answer generation and evidence extraction. This work represents an important step toward more reliable AI systems that make better use of available context without requiring expensive retrieval infrastructure.

References

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*. OpenReview.net, 2024.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., and Li, J. LongBench: A bilingual, multitask benchmark for long context understanding. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Chang, Y., Lo, K., Goyal, T., and Iyyer, M. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7Ttk3RzDeu.

Chen, M., Li, T., Sun, H., Zhou, Y., Zhu, C., Wang, H., Pan, J. Z., Zhang, W., Chen, H., Yang, F., et al. Research: Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.

Chen, Y., Chen, T., Jhamtani, H., Xia, P., Shin, R., Eisner, J., and Van Durme, B. Learning to retrieve iteratively for in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7156–7168, 2024.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. Deepseek-v3 technical report. *arXiv preprint arXiv: 2412.19437*, 2024.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv: 2501.12948*, 2025.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246/.

Farahani, M. and Johansson, R. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. *arXiv preprint arXiv:2410.05162*, 2024.

Fei, W., Niu, X., Xie, G., Zhang, Y., Bai, B., Deng, L., and Han, W. Retrieval meets reasoning: Dynamic in-context editing for long-text understanding. *arXiv preprint arXiv:2406.12331*, 2024.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020.

Ho, X., Duong Nguyen, A.-K., Sugawara, S., and Aizawa, A. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Scott, D., Bel, N., and Zong, C. (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Hsu, S., Khattab, O., Finn, C., and Sharma, A. Grounding by trying: Llms with reinforcement learning-enhanced retrieval. *arXiv preprint arXiv:2410.23214*, 2024.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, X., Ru, D., Qiu, L., Guo, Q., Zhang, T., Xu, Y., Luo, Y., Liu, P., Zhang, Y., and Zhang, Z. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv: 2405.14486*, 2024.

Humphreys, P., Guez, A., Tieleman, O., Sifre, L., Weber, T., and Lillicrap, T. Large-scale retrieval for reinforcement learning. *Advances in Neural Information Processing Systems*, 35:20092–20104, 2022.

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., and Han, J. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Kulkarni, M., Tangarajan, P., Kim, K., and Trivedi, A. Reinforcement learning for optimizing rag for domain chatbots. *arXiv preprint arXiv:2401.06800*, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

Li, H., Verga, P., Sen, P., Yang, B., Viswanathan, V., Lewis, P., Watanabe, T., and Su, Y. Alr^2: A retrieve-then-reason framework for long-context question answering. *arXiv preprint arXiv: 2410.03227*, 2024.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., Liu, H., Zhang, S., Song, K., Zhu, K., Cheng, Y., Wang, S., Wang, X., Luo, Y., Jin, H., Zhang, P., Liu, O., Chen, J., Zhang, H., Yu, Z., Shi, H., Li, B., Wu, D., Teng, F., Jia, X., Xu, J., Xiang, J., Lin, Y., Liu, T., Liu, T., Su, Y., Sun, H., Berseth, G., Nie, J., Foster, I., Ward, L., Wu, Q., Gu, Y., Zhuge, M., Tang, X., Wang, H., You, J., Wang, C., Pei, J., Yang, Q., Qi, X., and Wu, C. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv: 2504.01990*, 2025a.

Liu, S., Halder, K., Qi, Z., Xiao, W., Pappas, N., Htut, P. M., Anna John, N., Benajiba, Y., and Roth, D. Towards long context hallucination detection. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association* *for Computational Linguistics: NAACL 2025*, pp. 7827–7835, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.436/.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. Large language models: A survey. *arXiv preprint arXiv: 2402.06196*, 2024.

Nguyen, M., Nguyen, T. Q., KC, K., Zhang, Z., and Vu, T. Reinforcement learning from answer reranking feedback for retrieval-augmented answer generation. In *Proceedings of INTERSPEECH*, 2024.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. MS MARCO: A human generated machine reading comprehension dataset. In Besold, T. R., Bordes, A., d'Avila Garcez, A. S., and Wayne, G. (eds.), *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Post, M. A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K. (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y. Zero: Memory optimizations toward training trillion parameter models. *arXiv preprint arXiv: 1910.02054*, 2019.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. Hybridflow: A flexible and efficient rlhf framework. *European Conference on Computer Systems*, 2024. doi: 10.1145/3689031.3696075.

Song, H., Jiang, J., Min, Y., Chen, J., Chen, Z., Zhao, W. X., Fang, L., and Wen, J.-R. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025.

Talukdar, W. and Biswas, A. Improving large language model (llm) fidelity through context-aware grounding: A systematic approach to reliability and veracity. *arXiv preprint arXiv: 2408.04023*, 2024.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Tu, C.-H., Hsu, H.-J., and Chen, S.-W. Reinforcement learning for optimized information retrieval in llama. 2024.

Variengien, A. and Winsor, E. Look before you leap: A universal emergent decomposition of retrieval tasks in language models. *CoRR*, abs/2312.10091, 2023.

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J. W., Wei, J., Tar, C., Sung, Y., Zhou, D., Le, Q. V., and Luong, T. Freshllms: Refreshing large language models with search engine augmentation. In *ACL (Findings)*, pp. 13697–13720. Association for Computational Linguistics, 2024.

Wang, L., Yang, N., and Wei, F. Query2doc: Query expansion with large language models. In *EMNLP*, pp. 9414–9423. Association for Computational Linguistics, 2023.

Wang, L., Yang, N., and Wei, F. Learning to retrieve in-context examples for large language models. In *EACL (1)*, pp. 1752–1767. Association for Computational Linguistics, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wang, X., Chi, J., Tai, Z., Kwok, T. S. T., Li, M., Li, Z., He, H., Hua, Y., Lu, P., Wang, S., Wu, Y., Huang, J., Tian, J., and Zhou, L. Finsage: A multi-aspect rag system for financial filings question answering, 2025. URL https://arxiv.org/abs/2504.14493.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv: 1910.03771*, 2019.

Wu, J., Yang, L., Wang, Z., Okumura, M., and Zhang, Y. CofCA: A STEP-WISE counterfactual multi-hop QA benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=q2DmkZ1wVe.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. Benchmarking retrieval-augmented generation for medicine. In *ACL (Findings)*, pp. 6233–6251. Association for Computational Linguistics, 2024.

Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M., and Catanzaro, B. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Yan, S.-Q., Gu, J.-C., Zhu, Y., and Ling, Z.-H. Corrective retrieval augmented generation. 2024.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023a.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Yoran, O., Wolfson, T., Ram, O., and Berant, J. Making retrieval-augmented language models robust to irrelevant context. In *ICLR*. OpenReview.net, 2024.

Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., Desmaison, A., Balioglu, C., Damania, P., Nguyen, B., Chauhan, G., Hao, Y., Mathews, A., and Li, S. Pytorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860, 2023. doi: 10.14778/3611540.3611569. URL https://www.vldb.org/pvldb/vol16/p3848-huang.pdf.

Zheng, Y., Zhang, R., Zhang, J., Ye, Y., and Luo, Z. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In Cao, Y., Feng, Y., and Xiong, D. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.38. URL https://aclanthology.org/2024.acl-demos.38/.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

Zhou, Y., Dou, Z., and Wen, J.-R. Enhancing generative retrieval with reinforcement learning from relevance feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12481–12490, 2023.

Zhuang, S., Ma, X., Koopman, B., Lin, J., and Zuccon, G. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*, 2025.

A. Related Work

A.1. LLM Reasoning on Question-Answering Tasks

Large language models (LLMs) have demonstrated impressive capabilities in complex reasoning tasks (Wei et al., 2022; Cobbe et al., 2021; Ouyang et al., 2022). Recent work has explored various prompting strategies to improve reasoning, including chain of thought prompting (Wei et al., 2022), which guides models to generate intermediate reasoning steps before producing final answers, and its variants such as zero-shot-CoT (Kojima et al., 2022) and self-consistency (Wang et al., 2022). More structured approaches include tree-of-thought (Yao et al., 2023a), graph-of-thought (Besta et al., 2024), ReAct (Yao et al., 2023b), and least-to-most prompting (Zhou et al., 2022). Despite these advances, LLMs still struggle with maintaining context coherence when reasoning over long or noisy inputs (Xu et al., 2023; Li et al., 2024; Fei et al., 2024).

A.2. Retrieval-Augmented Generation

Traditional retrieval-augmented generation (RAG) methods (Guu et al., 2020; Lewis et al., 2020) enhance LLM by retrieving relevant passages from external corpora, alleviating the limitations of fixed parametric memory. This framework has been widely adopted for knowledge-intensive tasks (Xiong et al., 2024; Wang et al., 2025). Recent work has improved retrieval quality through techniques such as query expansion (Wang et al., 2023), re-ranking (Vu et al., 2024), and filtering (Asai et al., 2024), while others focus on robustness to noisy retrievals (Yoran et al., 2024). In-context retrieval methods aim to reuse relevant spans from the input sequence itself (Variengien & Winsor, 2023; Wang et al., 2024). However, both external and in-context RAG fundamentally rely on indexing and embedding-based retrieval pipelines, limiting their adaptability to complex or evolving contexts.

A.3. RL-Enhanced LLM Retrieval

Reinforcement learning (RL) has emerged as a powerful paradigm for optimizing LLM retrieval strategies (Humphreys et al., 2022; Tu et al., 2024; Hsu et al., 2024). Unlike traditional retrieval methods, RL-based approaches can learn adaptive retrieval policies that optimize for task-specific rewards (Kulkarni et al., 2024; Zhuang et al., 2025; Jin et al., 2025). Recent work has explored using RL to train retrieval policies that maximize answer correctness (Hsu et al., 2024; Nguyen et al., 2024), combining the strengths of parametric knowledge and nonparametric retrieval (Mallen et al., 2022; Humphreys et al., 2022; Farahani & Johansson, 2024). Several approaches have used feedback mechanisms to improve retrieval quality, including relevance feedback (Zhou et al., 2023) and iterative refinement (Chen et al., 2024). However, most existing approaches still maintain a separation between the retrieval mechanism and the core reasoning process, potentially limiting the model's ability to integrate retrieved information in a context-aware manner.

B. The `CARE` RL Training Algorithm

Below, we present the RL training algorithm with curriculum learning of `CARE` in Algorithm 1.

C. Experiment Settings

C.1. Datasets, Benchmarks and Metrics

Training Datasets. We generate the SFT data mentioned in Section 2.2 based on the HotpotQA training set (Yang et al., 2018) thanks to its supporting facts annotations. During SFT data generation, DeepSeek-R1 (DeepSeek-AI et al., 2025) and DeepSeek-V3 (DeepSeek-AI et al., 2024) are used as the reasoning model M_R and the fact injection model M_I, respectively. The resulting SFT dataset contains 7,739 instances with the retrieval-augmented reasoning chain labeled. For RL training, we select DROP (Dua et al., 2019) as $\mathcal{D}_{\text{easy}}$ and MS MARCO (Nguyen et al., 2016) as $\mathcal{D}_{\text{hard}}$.

Evaluation Datasets. We assess in-context retrieval accuracy and whether learned retrieval-augmented reasoning improves answer quality using both single-passage and multi-passage datasets from LongBench (Bai et al., 2024), including **MultiFieldQA-En** (Bai et al., 2024), **HotpotQA** (Yang et al., 2018), **2WikiMQA** (Ho et al., 2020), and **MuSiQue** (Trivedi et al., 2022). Following LongBench's protocol, we report F1 scores for all datasets.

Furthermore, to evaluate context fidelity when presented with information contradicting the model's parametric knowledge, we utilize **CofCA** (Wu et al., 2025), a benchmark containing modified counterfactual Wikipedia snippets. This directly tests

Algorithm 1 Curriculum RL with CARE Rewards

Require: Datasets $\mathcal{D}_{\text{easy}}$, $\mathcal{D}_{\text{hard}}$, policy π_θ, reference policy π_{ref}, clip range ϵ, KL coefficient β, initial ratio $\alpha = 1.0$, total steps T
Ensure: Updated policy parameters θ
for each training step t **do**
  Sample query q with probability α from $\mathcal{D}_{\text{easy}}$ and $1 - \alpha$ from $\mathcal{D}_{\text{hard}}$
  Sample outputs $\{o_i\}_{i=1}^{G}$ from $\pi_{\theta_{\text{old}}}(q)$
  for each output o_i **do**
    Extract retrieval spans S from o_i
    Compute the Retrieval, Format and Accuracy Rewards defined in Section 2.3
    for each token t in o_i **do**
      Compute importance ratio $r_{i,t} = \frac{\pi_\theta(o_{i,t})}{\pi_{\theta_{old}}(o_{i,t})}$
      Update objective with Equation 1
    end for
  end for
  Apply KL penalty: $J_{\text{GRPO}} \leftarrow J_{\text{GRPO}} - \beta \sum_t \pi_\theta(o_t) \log\left(\frac{\pi_\theta(o_t)}{\pi_{\text{ref}}(o_t)}\right)$
  Update parameters: $\theta \leftarrow \theta + \eta \nabla_\theta J_{\text{GRPO}}$
  Adjust curriculum ratio $\alpha \leftarrow \max(0, 1 - \eta t/T)$
end for
return θ

whether our native retrieval-augmented reasoning improves adherence to provided context regardless of pre-trained biases. We report F1 performance consistent with CofCA's original evaluation metrics.

C.2. Models and Baselines

We compare CARE with a series of learned reasoning strategies and RAG methods based on three commonly used public LLMs: Qwen-2.5 7B and 14B and LLaMA-3.1 8B, which covers different model families and sizes.

Original Model. For each dataset, we test the performance of the original LLM with their corresponding default system prompt and chat template.

RL-Based Online Retrieval. Existing dynamic retrieval approaches typically leverage reinforcement learning to train models to autonomously conduct web searches rather than directly extract from provided context. We compare our method against two recent RL-based online search methods: ReSearch (Chen et al., 2025) and R1-Searcher (Song et al., 2025), both of which enable models to strategically access external knowledge during reasoning. Note that in our model selection, ReSearch only provides a checkpoint for Qwen2.5 7B, and R1-Searcher only provides checkpoint for LLaMA-3.1 8B and Qwen2.5 7B.

RAG Methods. We also compare with CRAG (Yan et al., 2024), a corrective RAG method that uses a lightweight evaluator to enhance in-context retrieval with online searching. Note that in our model selection, CRAG only provides a checkpoint for Qwen2.5 7B and 14B.

C.3. Implementation Details

All models are implemented based on the pretrained checkpoints provided by the Huggingface Transformers library (Wolf et al., 2019). We use LLaMA-Factory (Zheng et al., 2024) for the SFT phase. In this phase, we train each model on our curated SFT dataset for 3 epochs with the AdamW optimizer (Loshchilov & Hutter, 2019). The training progress adopts a warmup cosine scheduler with maximum learning rate 0.0001 and warmup ratio 0.1. The effective batch size is 64. LoRA (Hu et al., 2022) is applied with $r = 8$ and $\alpha = 16$. The training process uses ZeRO-2 optimizer (Rajbhandari et al., 2019). For the RL phase, we adopt the verl framework (Sheng et al., 2024) for GRPO training. We used a training batch size of 1024. The Adam optimizer was employed with a learning rate of 1e-6. For policy optimization, we use GRPO as the advantage estimator, and incorporated KL divergence regularization with a coefficient of 0.001 using the low-variance KL estimator. We set the mini-batch size to 256. The model was trained for 350 steps with 5 response samples per prompt. For distributed training, we deployed Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023) across 8 GPUs on a single node with tensor parallelism of size 2. All experiments are done with either 8×A800-SXM4-80GB or 8×H100 80GB.

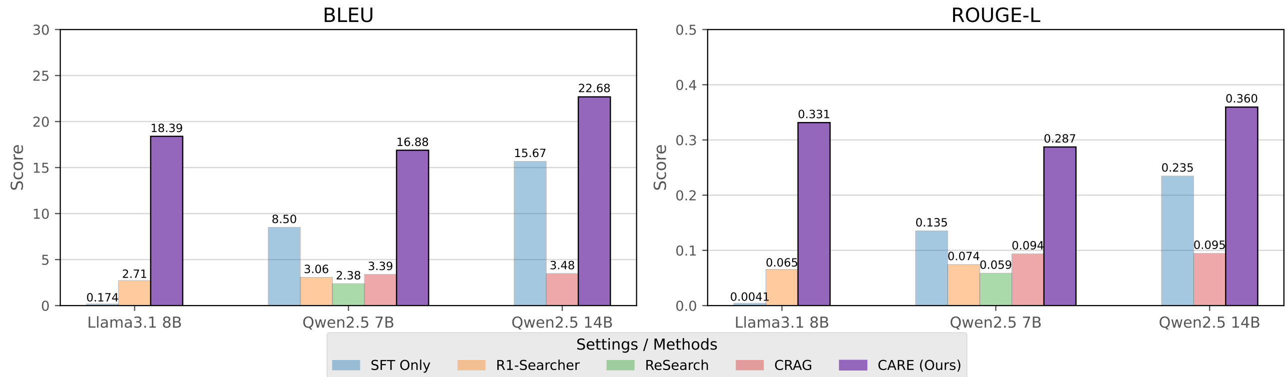

Figure 2. Comparison of models' retrieval accuracy across different settings in terms of BLEU and ROUGE-L metrics. Our proposed methods, `CARE`, demonstrate improved scores.

D. Evidence Retrieval Evaluation

In this section, we evaluate `CARE`'s ability to accurately retrieve and incorporate supporting evidence for question-answering. Due to the lack of ground-truth supporting fact annotations in standard QA datasets, we focus our evaluation on the LongBench HotpotQA benchmark. For this analysis, we align each instance in LongBench's HotpotQA test set with its corresponding entry in the original HotpotQA dataset, using the original supporting fact annotations as ground truth for evaluation. We report SacreBLEU (Post, 2018) and ROUGE-L F1 (Lin, 2004). Figure 2 presents our comparative results in different model configurations. Across all settings, `CARE` consistently achieves the highest BLEU and ROUGE-L scores. We observe that performance scales with model size across all methods, with Qwen2.5 14B showing the strongest results. However, the relative improvement from `CARE` remains consistent regardless of model scale and family, suggesting that our approach effectively enhances context fidelity regardless of underlying model architecture.

E. System Prompts

We provide the system prompts used in the dataset creation process and the `CARE` below.

Prompt used for M_R's generation of reasoning chains for SFT data creation.

You're an expert reader. Your goal is to read a context to answer a question. Note that during your thinking process, before you make *any reasoning step that requires retrieving information from the context*, summarize what information you would need to complete this reasoning step, such as "I need to know X for this" or similar phrases before you reason about the context. This will help you to be more systematic in your reasoning process. Put your final answer as a minimum phrase or word at the end after "Answer:".
Context: {context}
Question: {question}

Prompt used for M_I's evidence integration for SFT data creation.

I'll provide you with a question, a reasoning process to solve this question, and several evidence sentences. Insert *all* evidence sentences into the reasoning process at appropriate locations and give me the updated reasoning process. Each evidence sentence usually should be placed just before any conclusions or deductions that depend on it. The evidence sentences may need to be distributed throughout different parts of the reasoning and may appear more than once. *Do not modify any evidence sentences* - insert them exactly as provided. Return only the completed reasoning process without explanations or additional text scaffolds.
Question: {question}
Reasoning process: {reasoning_content}

Evidence sentences (One sentence per line): {evidence_sentence_string}
The rewritten reasoning process:

System prompt for CARE. The actual system prompt for each model prepends the corresponding model's original system prompt before this prompt.

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within ⟨ think ⟩ ⟨ /think ⟩ tags. WITHIN the thinking process, make reference to the relevant texts in the prompt that provide critical information to move the reasoning process forward. The referenced texts MUST BE enclosed within ⟨ retrieval ⟩ ⟨ /retrieval ⟩ tags, and MUST BE placed within the reasoning process only. The final answer MUST BE put at the end of the response after "Answer:".

F. Case Study

We select two samples from LongBench's HotpotQA dataset and provide the complete question, answer, supporting facts, prediction, and whether the generated answer is correct.

Question	Answer	Supporting Facts	Prediction	Correct
The song that came out ahead of The Wanted's “Lightning” on the UK Singles Chart was featured on which album?	Talk That Talk	["“We Found Love” is a song recorded by Barbadian singer Rihanna from her sixth studio album, “Talk That Talk” (2011).", "It was written and produced by, and features Scottish DJ Calvin Harris; the song was also featured on his third album “18 Months” (2012).", "“We Found Love” premiered on September 22, 2011, on the Capital FM radio station in the United Kingdom, and was made available to download on the same day as the lead single from “Talk That Talk”.", "“We Found Love” is an uptempo electro house song, with elements of techno and Europop.", "The song's lyrics speak of a couple who “found love in a hopeless place”."]	⟨ think ⟩\nOkay, let's try to figure out the answer to this question. The question is asking which album featured the song that came out ahead of The Wanted's “Lightning” on the UK Singles Chart.\n\nFirst, I need to find out the details about The Wanted's “Lightning” from the passages provided. Looking at the passage for The Wanted's discography, it mentions that their third single from the second album “Battleground” was “Lightning,” released in October 2011. ⟨ retrieval ⟩The song debuted at number two on the UK Singles Chart, behind Rihanna's “We Found Love”.⟨ /retrieval ⟩ So, that means the song that was number one before “Lightning” was “We Found Love” by Rihanna.\n\nNow, the question is about the album that features “We Found Love.” I need to check where “We Found Love” is mentioned. In the passages, there's a section about Rihanna's “We Found Love” from her album “Talk That Talk (2011.” The album is listed as her sixth studio album. The song itself is part of that album. ⟨ retrieval ⟩“We Found Love” is a song by Barbadian singer Rihanna from her sixth studio album, Talk That Talk (2011.⟨ /retrieval ⟩ So the answer should be “Talk That Talk.”\n\nWait, but let me confirm. The relevant information from the third passage (The Rihanna section for “We Found Love” clearly states that it's from “Talk That Talk (2011),” and that's her sixth studio album. The song's debut was with Rihanna's album. Therefore, the answer is “Talk That Talk.”\n⟨ /think ⟩\n\nAnswer: Talk That Talk	Yes
Yau Ma Tei North is a district of a city with how many citizens?	7.2 million	["Hong Kong, officially the Hong Kong Special Administrative Region of the People's Republic of China, is an autonomous territory south to Mainland China and east to Macao in East Asia.", " With around 7.2 million Hong Kongers of various nationalities in a territory of 1,104 km, Hong Kong is the world's fourth most densely populated country or territory."]	⟨ think ⟩\nOkay, let's tackle this question. The question is asking how many citizens Yau Ma Tei North is a district of. First, I need to figure out which city Yau Ma Tei North is a district of. From the context, the term “Yau Ma Tei North” is mentioned in several passages. Let me check.\n\nLooking at the passages, the one titled “Yau Ma Tei North (constituency)” states that Yau Ma Tei North is one of the 19 constituencies in the Yau Tsim Mong District of Hong Kong. \n⟨ retrieval ⟩Yau Ma Tei North is one of the 19 constituencies in the Yau Tsim Mong District of Hong Kong which was first created in 1982 and recreated in 2015.⟨ /retrieval ⟩\nThat's in the context of Hong Kong's administrative divisions. The next step would be to find out the population of Hong Kong. \n\nI recall that there's a detailed section in the “Hong Kong” passages about the population. Let me check that part. The Hong Kong entry mentions that the Census and Statistics Department estimated Hong Kong's population at 7,413,070 in 2021.\n⟨ retrieval ⟩The Census and Statistics Department estimated Hong Kong's population at 7,413,070 in 2021. The overwhelming majority (91.6%) is Han Chinese, most of whom are Taishanese, Teochew, Hakka, and other Cantonese peoples.⟨ /retrieval ⟩\n\nWait, but the question is about Yau Ma Tei North, which is a district within Hong Kong. The population of Hong Kong as a whole is 7.4 million, but the key point here is whether Yau Ma Tei North itself has a population figure. The Yau Ma Tei North constituency's entry says it covers Yau Ma Tei and has an estimated population of 12,817. \n⟨ retrieval ⟩The constituency loosely covers Yau Ma Tei with the estimated population of 12,817.⟨ /retrieval ⟩\nSo that's 12,817 people. But the question is phrased as “Yau Ma Tei North is a district of a city with how many citizens?” So the target is the parent city of Yau Ma Tei North. Since Yau Ma Tei North is a district in Hong Kong, Hong Kong is the city. Therefore, the population of Hong Kong is the answer, which is 7,413,070 as per the latest data provided.\n⟨ /think ⟩\n\nTo determine the population of the city (Hong Kong) containing Yau Ma Tei North, I need to know the administrative level of Yau Ma Tei North. The “Yau Ma Tei North (constituency)” section states it belongs to Yau Tsim Mong District. Next, the “Hong Kong” passages mention the district is part of the Yau Tsim Mong District and its population. The main Hong Kong entry (under “Demographics”) gives the 2021 population as 7,413,070. However, Yau Ma Tei North is a specific constituency (the 19th out of 19) within Yau Tsim Mong District, with an estimated population of 12,817. The question likely refers to the broader city of Hong Kong's population, not the smaller constituency. \nAnswer: 7413070	No