distillation of query rewriting is sub-optimal.

The scores on multiple-choice QA are presented in Table 3. With ChatGPT as a reader, it can be observed that query rewriting improves the scores in most of the settings, except for the social sciences category. With Vicuna as a reader, our method achieves more gains on the four categories compared to ChatGPT. This agrees with the intuition that a more powerful reader has more parametric memories, thus more difficult to compensate with external knowledge.

Model	**EM**	**F_1**
HotpotQA		
Direct	32.36	43.05
Retrieve-then-read	30.47	41.34
LLM rewriter	32.80	43.85
Trainable rewriter	34.38	45.97
AmbigNQ		
Direct	42.10	53.05
Retrieve-then-read	45.80	58.50
LLM rewriter	46.40	58.74
Trainable rewriter	47.80	60.71
PopQA		
Direct	41.94	44.61
Retrieve-then-read	43.20	47.53
LLM rewriter	46.00	49.74
Trainable rewriter	45.72	49.51

Table 2: Metrics of open-domain QA.

MMLU	**EM**			
	Human.	STEM	Other	Social
ChatGPT				
Direct	75.6	58.8	69.0	71.6
Retrieve-then-read	76.7	63.3	70.0	78.2
LLM rewriter	77.0	63.5	72.6	76.4
Vicuna-13B				
Direct	39.8	34.9	50.2	46.6
Retrieve-then-read	40.2	39.8	55.2	50.6
LLM rewriter	42.0	41.5	57.1	52.2
Trainable rewriter	43.2	40.9	59.3	51.2

Table 3: Metrics of multiple choice QA.

6 Analysis

6.1 Training Process

The training process includes two stages, warm-up and reinforcement learning. This section shows the validation scores of the three open-domain QA datasets for further analysis. Figure 2 presents the metric scores through training iterations in the process of reinforcement learning. As the rewriting models have been warmed up on the pseudo data before RL, scores at "0 iteration" denote the ability acquired from the warm-up training.

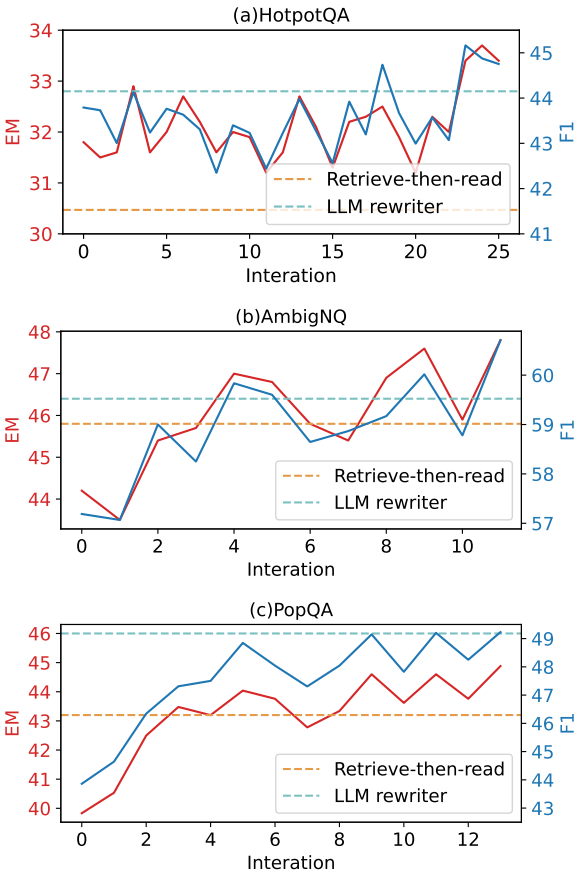

Figure 2: Reinforcement learning validation scores of (a)HotpotQA, (b)AmbigNQ, and (c)PopQA. The solid lines show EM (red) and F1 (blue) numbers through training iterations. The dashed lines are EM scores of the standard retrieve-then-read method (orange) and retrieval with an LLM as the rewriter (green).

It can be observed that the curves show upward trends with some fluctuations on all the datasets. (i) For multi-hop questions in HotpotQA, the standard retrieval is relatively weaker. Complex questions can be not specific search queries and show a larger gap from rewritten queries, i.e., the green and red lines. (ii) On AmbigNQ and PopQA, our method surpasses the baselines after several iterations (3 or 4). This indicates that the RL training stage can compensate for the insufficiency of the distillation on the pseudo data during warm-up training. (iii) In particular, on PopQA, the trainable rewriter remains inferior to the LLM rewriter. This can be explained as the dataset is constructed for adaptive retrieval (Mallen et al., 2022), which only uses retrieval where it helps to avoid harmful redundant retrieval. Thus, *"None"* is a possible query that means no retrieval. This causes more complexity and uncertainty. LLM rewriter knows better when the retrieval is needed for itself as a reader, although the rewriting step is not concatenated as

the input context of the reader.

We calculate the performance of query *"None"*. The questions that can be correctly answered without retrieval (i.e., the "Direct" method) are those samples that need no more context. Comparing this retrieval-free set with those that are rewritten to be *"None"* query, the F_1 score of the LLM rewriter is 71.9% and the T5 rewriter score is 67.1%. If we consider the questions that can be correctly answered without retrieval but go wrong with retrieval as the retrieval-free set, the F_1 scores are 78.7% for LLM rewriter and 77.4% for T5.

Model	EM	F_1	Hit ratio
No retrieval	42.10	53.05	–
Upper bound	58.40	69.45	100
Retrieve-then-read			
w/ snippet	38.70	50.50	61.1
w/ BM25	45.80	58.50	76.4
LLM rewriter			
w/ snippet	39.80	52.64	63.5
w/ BM25	46.40	58.74	77.5
Trainable rewriter			
w/ BM25[2]	47.80	60.71	82.2

Table 4: Retrieval analysis on AmbigNQ.

6.2 Retrieval Result

Our proposed method is a pipeline framework, instead of an end-to-end system. The query rewriting first affects the retrieved context, then the context makes a difference to the output of the reader. Hence, QA metrics are indirect measurements. We take a closer look at the retrieved context and the reader capability through the retrieval metric, hit ratio. After text normalization, the hit rate is computed to measure whether the retrieved context contains the correct answers.

Table 4 shows the scores on AmbigNQ. The scores in the second line are computed on a selection of the samples whose retrieved contexts hit correct answers (under the standard retrieve-then-read setting). The scores show the approximate upper bound ability of the reader with retrieval augmentation, abbreviated as the "upper bound" score. The effectiveness of retrieval is proved compared to the no retrieval setting (the first line). For each retrieval method, two settings are presented: (i) collecting Bing snippets, (ii) selecting from URLs by BM25. The metrics show that content selection with BM25 recalls better documents than snippets, while query rewriting makes progress on both settings. We also observed that the improvement in the hit rate of the retriever is more significant than the improvement in the reader. This is consistent with the findings in related search (Mallen et al., 2022; Liu et al., 2023).

[2]Our trainable rewriter is adapted to the retriever using BM25 during RL training. Using the output queries of the test set after training, the snippet hit rate is 73.4%.

Example 1: multi-hop question	Hit	Correct
Q0: The youngest daughter of Lady Mary-Gaye Curzon stars with Douglas Smith and Lucien Laviscount in what 2017 film?	✗	✗
Q1: the youngest daughter of Lady Mary-Gaye Curzon; 2017 film stars Douglas Smith and Lucien Laviscount	✓	✓
Q2: Lady Mary-Gaye Curzon youngest daughter 2017 film with Douglas Smith and Lucien Laviscount	✓	✓
Example 2:		
Q0: What 2000 movie does the song "All Star" appear in?	✗	✗
Q1: movie "All Star" 2000	✗	✗
Q2: 2000 movie "All Star" song	✓	✓
Example 3: multiple choice		
Q0: A car-manufacturing factory is considering a new site for its next plant. Which of the following would community planners be most concerned with before allowing the plant to be built? Options: A. The amount of materials stored in the plant B. The hours of operations of the new plant C. The effect the plant will have on the environment D. The work environment for the employees at the plant	✗	✗
Q1: What would community planners be most concerned with before allowing a car-manufacturing factory to be built?	✓	✓

Figure 3: Examples for intuitive illustration. Q0 denotes original input, Q1 is from the LLM rewriter, and Q2 is from the trained T5 rewriter. **Hit** means retriever recall the answer, while **Correct** is for the reader output.

6.3 Case Study

To intuitively show how the query rewriting makes a difference in the retrieved contexts and prediction performance, we present examples in Figure 3 to compare the original questions and the queries. In example 1, the original question asks for a film that *the youngest daughter of Lady Mary-Gaye Curzon* co-stars with two certain actors. Both query 1 and query 2 put the keyword *film* forward, closely following *the youngest daughter of Lady Mary-Gaye Curzon*. With both, the actress *Charlotte Calthorpe* and her movie information can be retrieved and the answer is included. The second is an example where the query from the LLM rewriter failed but

the query from T5 gets the correct answer. The number *2000* is misunderstood in query 1, while query 2 keeps *200 movie* together, avoiding meaningless retrieval. Example 3 is for multiple choice. The query simplifies the background and enhances the keyword *community planner*. The retrieve contexts are mainly about *Introduction to Community Planning* where the answer *environment* appears several times.

7 Conclusion

This paper introduces the *Rewrite-Retrieve-Read* pipeline, where a query rewriting step is added for the retrieval-augmented LLM. This approach is applicable for adopting a frozen large language model as the reader and a real-time web search engine as the retriever. Further, we propose to apply a tuneable small language model the rewriter, which can be trained to cater to the frozen retriever and reader. The training implementation consists of two stages, warm-up and reinforcement learning. Evaluation and analyses on open-domain QA and multiple-choice QA show the effectiveness of query rewriting. Our work proposes a novel retrieval-augmented black-box LLM framework, proves that the retrieval augmentation can be enhanced from the aspect of query rewriting, and provides a new method for integrating trainable modules into black-box LLMs.

Limitations

We acknowledge the limitations of this work. (i) There is still a trade-off between generalization and specialization among downstream tasks. Adding a training process, the scalability to direct transfer is compromised, compared to few-shot in-context learning. (ii) The research line of *LLM agent* has shown impressive performance but relies on multiple calls to the LLM for each sample (Khattab et al., 2022; Yao et al., 2023), where the LLM plays as an agent to flexibly call the retriever multiple times, reads the context in earlier hops, and generates follow-up questions. Different from these studies, our motivation is to enhance the one-turn retriever-then-read framework with a trainable query rewriter. (iii) Using a web search engine as the retriever also leads to some limitations. Neural dense retrievers that are based on professional, filtered knowledge bases may potentially achieve better and controllable retrieval. More discussion is included in the appendix.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2022. Can retriever-augmented language models reason? the blame game between the retriever and the language model. *arXiv preprint arXiv:2212.09146*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of