RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering

Zihan Zhang[1], Meng Fang[2], Ling Chen[1]
[1]University of Technology Sydney [2]University of Liverpool
Zihan.Zhang-5@student.uts.edu.au, Meng.Fang@liverpool.ac.uk
Ling.Chen@uts.edu.au

Abstract

Adaptive retrieval-augmented generation (ARAG) aims to dynamically determine the necessity of retrieval for queries instead of retrieving indiscriminately to enhance the efficiency and relevance of the sourced information. However, previous works largely overlook the evaluation of ARAG approaches, leading to their effectiveness being understudied. This work presents a benchmark, `RetrievalQA`, comprising 1,271 short-form questions covering new world and long-tail knowledge. The knowledge necessary to answer the questions is absent from LLMs; therefore, external information must be retrieved to answer correctly. This makes `RetrievalQA` a suitable testbed to evaluate existing ARAG methods. We observe that calibration-based methods heavily rely on threshold tuning, while vanilla prompting is inadequate for guiding LLMs to make reliable retrieval decisions. Based on our findings, we propose **T**ime-**A**ware **A**daptive **RE**trieval (**TA-ARE**), a simple yet effective method that helps LLMs assess the necessity of retrieval without calibration or additional training[1].

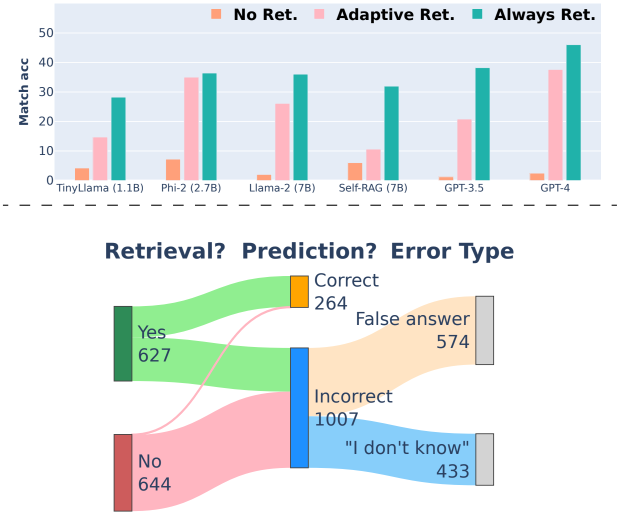

Figure 1: **Above**: QA accuracy on our `RetrievalQA` w/, w/o retrieval, and adaptive retrieval. We set threshold $t = 0.5$ for *calibration-based* Self-RAG (Asai et al., 2023b) and use *model-based* **Vanilla** prompting for others (§3). We find that Self-RAG requires threshold tuning to balance QA performance and retrieval efficiency, while vanilla prompting is insufficient in guiding LLMs to make reliable retrieval decisions (§3.3). **Below**: an error analysis for GPT-3.5. At least half of the time, GPT-3.5 is unaware that it needs retrieval (*i.e.*, Red area, §3.4).

1 Introduction

Retrieval-augmented generation (RAG) (Guu et al., 2020; Lewis et al., 2020; Ram et al., 2023) that augments large language models (LLMs) with retrieval of relevant information has become increasingly popular in knowledge-intensive tasks, including open-domain question-answering (QA) (Zhang et al., 2023b; Kasai et al., 2023; Cui et al., 2023; Zhang et al., 2023a). However, standard RAG methods conduct retrieval *indiscriminately*, irrespective of the input query, which may result in suboptimal task performance and increased inference costs (Gao et al., 2024). On one hand, LLMs encode vast knowledge in parameters through large-scale pre-training, enabling them to effortlessly handle straightforward queries without retrieval (Mallen et al., 2023). On the other hand, the retrieved context may contain noise and irrelevant information, and augmenting noisy context can potentially distract LLMs, thereby impeding task performance (Shi et al., 2023).

To alleviate the limitations of RAG mentioned above, recent studies advocate for **adaptive RAG (ARAG)**, which dynamically determines retrieval necessity and relies only on LLMs' parametric knowledge when deemed unnecessary (Feng et al., 2023b). However, the effectiveness of these meth-

[1]The dataset and code are available at `https://github.com/hyintell/RetrievalQA`

Findings of the Association for Computational Linguistics: ACL 2024, pages 6963–6975

ods is understudied, as there is no suitable benchmark and evaluation. ARAG approaches can be categorized into *calibration-based* and *model-based* judgement. Calibration-based methods (Mallen et al., 2023; Jiang et al., 2023; Asai et al., 2023b), while effective, trigger retrieval only when a metric surpasses a pre-defined threshold. For example, Mallen et al. (2023) heuristically retrieve when the popularity of an entity on Wikipedia is below a certain threshold; Jiang et al. (2023) trigger retrieval if any token in the temporarily generated sentence has low confidence. Clearly, these ad-hoc calibration-based methods are suboptimal, as we need to tune thresholds for different datasets and models to balance task performance and inference overheads. To obviate the hyperparameter threshold, model-based methods (Feng et al., 2023a; Ren et al., 2023) directly prompt LLMs for retrieval decisions, given the observation that LLMs can acknowledge their knowledge boundaries to some extent (Kadavath et al., 2022; Yin et al., 2023). These methods undergo separate evaluations, and their effectiveness remains ambiguous due to the limited scope of the assessments.

In this paper, we investigate to what extent LLMs can perform calibration-free adaptive retrieval via prompting. To answer this question, we need to evaluate whether LLMs retrieve *only* when necessary. This requests a benchmark that distinguishes between questions that can be answered using LLMs' parametric knowledge and those that require external information through retrieval. Nevertheless, commonly used open-domain QA datasets (Rajpurkar et al., 2016; Joshi et al., 2017; Kwiatkowski et al., 2019; Mallen et al., 2023) fail to fulfil this purpose, as various LLMs have distinct sizes and levels of pre-trained knowledge, making them inadequately assess the necessity of external retrieval for LLMs.

To fill this gap, we create **RetrievalQA**, a short-form QA dataset, covering new world and long-tail knowledge and spanning diverse topics. We ensure the knowledge necessary to answer the questions is absent from LLMs. Therefore, LLMs must truthfully decide whether to retrieve to be able to answer the questions correctly. RetrievalQA enables us to evaluate the effectiveness of ARAG approaches, an aspect predominantly overlooked in prior studies and recent RAG evaluation systems (Chen et al., 2023; Saad-Falcon et al., 2023; Es et al., 2023), which focus only on task performance, the relevance of retrieval context or the faithfulness of answers.

Using RetrievalQA as a testbed, we benchmark both calibration-based and model-based methods with varying sizes of LLMs. As shown in Fig.1, we find calibration-based Self-RAG requires threshold tuning to balance QA performance and retrieval efficiency, while vanilla prompting is insufficient in guiding LLMs to make reliable retrieval decisions. As an initial effort, we propose **T**ime-**A**ware **A**daptive **RE**trieval (**TA-ARE**), a simple yet effective method to improve ARAG via in-context learning (ICL; Brown et al. 2020), obviating the need for calibration or additional training.

To sum up, this paper makes the following contributions: ① we create a new dataset RetrievalQA to assess ARAG for short-form open-domain QA; ② we benchmark existing methods and conduct extensive analysis, finding that vanilla prompting is insufficient in guiding LLMs to make reliable retrieval decisions; ③ we then propose TA-ARE, a simple yet effective method to help LLMs assess the necessity of retrieval without calibration or additional training.

2 Dataset Construction

Data collection. Inspired by Zhuang et al. (2023), we aim to collect data such that the knowledge necessary to answer the questions is absent from LLMs. Therefore, LLMs must consult external resources to answer correctly. Specifically, we mainly collect data from two categories: ① **New world knowledge** that is out of the scope of the LLMs' pre-training corpora. LLMs are static after training and can quickly be outdated due to the ever-changing world (Zhang et al., 2023b). To ensure the knowledge is novel to most LLMs, we select 397 QA pairs ranging from 1 October 2023 to 12 January 2024 from RealTimeQA (Kasai et al., 2023). These data comprise weekly quizzes extracted from news websites, encompassing broad topics, including politics, business, and entertainment. In addition, we collect 127 fast-changing questions from FreshQA (Vu et al., 2023), where the answers may change frequently, thereby challenging LLMs' parametric memorization. ② **Long-tail knowledge** that is rarely learned during pre-training. Previous studies (Kandpal et al., 2023) have shown that LLMs struggle to learn less common knowledge and perform poorly without the help of retrieval. Following Asai et al. (2023b), we use the long-tail subset of PopQA (Mallen et al., 2023), which consists of 1,399 rare entity queries with monthly Wikipedia page views below 100, and the test split of unfiltered TriviaQA (Joshi et al., 2017), which has 7,313 factual QA pairs. Lastly, we collect 100 personal agenda questions from ToolQA (Zhuang et al., 2023), which are synthesized with virtual names and events.

Filtering out answerable questions. As discussed in §1, to ensure the questions cannot be answered without external knowledge, we conduct strict filtering. To save manual work, we prompt GPT-4 for answers in a closed-book QA setting without access to external knowledge (see prompt template Fig.7 in Appendix). Then, we calculate the token-level F1 scores (Rajpurkar et al., 2016) and remove questions that have shared tokens between the prediction and the ground truth, *i.e.*, only keep questions with F1 $= 0$. Our rationale is that if state-of-the-art GPT-4 cannot answer correctly without retrieval, weaker LLMs are also highly likely to fail. Finally, after filtering, we have obtained 1,271 out of 9,336 questions, covering new world and long-tail knowledge and spanning diverse topics. To avoid potential bias in the evaluation towards methods that retrieve more often, we additionally collect 1,514 questions that can be answered using GPT-2's parametric knowledge from the discard set. More details are in the Appendix A.2.

We conduct a sanity check using various sizes of LLMs in Fig.1 and in Appendix A.5, showing that `RetrievalQA` is extremely hard for all models without access to external knowledge. We present detailed data statistics in Table 3 and examples of the data in Table 11.

3 `RetrievalQA` Challenges Adaptive RAG

In this section, we formalize ARAG for open-domain QA tasks and evaluate existing adaptive approaches on `RetrievalQA`.

3.1 Standard & Adaptive RAG for QA

Standard RAG. Given a question x, a retriever $\mathcal{R}$, and an external document corpus $\mathcal{D}$ such as Wikipedia, the retriever first retrieves a list of relevant documents $\mathcal{D}_x = \mathcal{R}(x)$, then an LLM needs to generate answer $y = \text{LLM}(I, \mathcal{D}_x, x)$ conditioned on a prompt instruction I, retrieved documents $\mathcal{D}_x$, and the question x.

Adaptive RAG. Standard RAG always retrieves regardless of the input question, while adaptive retrieval only retrieves when necessary. **Calibration-based** methods generally introduce a pre-defined hyperparameter t and only do retrieval when a metric surpasses t:

$$y = \begin{cases} \text{LLM}(I, \mathcal{D}_x, x), & \text{metric } \geq t \\ \text{LLM}(I, x), & \text{otherwise} \end{cases}$$

Baselines. While our primary focus is model-based methods, we also evaluate the most recent state-of-the-art Self-RAG (7B) (Asai et al., 2023b). Self-RAG fine-tunes Llama-2 using special reflection tokens to allow the model to introspect its outputs. The model activates retrieval when the probabilities of the generated special tokens exceed a threshold.

For **Model-based** methods, we follow Feng et al. (2023a) and Ren et al. (2023) to instruct LLMs to decide whether to retrieve via prompting, obviating the threshold. Specifically, we ask a yes/no question: $r = \text{LLM}(I_{\text{vanilla}}, x)$, where $I_{\text{vanilla}} =$ `"Given a question, determine whether you need to retrieve ... answer [Yes] or [No]"`. Retrieval is performed only when LLMs answer yes. We denote this as **Vanilla** prompting:

$$y = \begin{cases} \text{LLM}(I, \mathcal{D}_x, x), & r = \text{Yes} \\ \text{LLM}(I, x), & \text{otherwise} \end{cases}$$

Baselines. We evaluate strong instruction-tuned models with a varying scale of model size: TinyLlama (1.1B; Zhang et al. 2024), Phi-2 (2.7B; Li et al. 2023), Llama-2 (7B; Touvron et al. 2023), GPT-3.5 (OpenAI, 2022), and GPT-4 (OpenAI, 2023).

3.2 Experiment Setup

Evaluation Metric. We use *retrieval* accuracy to evaluate how well LLMs can perform adaptive retrieval. Since all questions in our dataset need retrieval, the higher the retrieval accuracy, the more effective the method. Following Schick et al. (2023); Mallen et al. (2023); Asai et al. (2023b), we evaluate QA performance using *match* accuracy, which measures whether gold answers are included in the model predictions instead of strict exact matching.

Implementation details. For Self-RAG, we set the retrieval threshold $t = [0.25, 0.5, 0.75, \texttt{None}]$. Lower thresholds encourage more frequent retrieval, while `None` means the model itself decides when to retrieve by generating the specific `[Retrieval]` token. Since the quality of the retrieved documents is not the focus of this paper, for long-tail knowledge questions, we use

Baselines (1,271)	Adaptive Retrieval		Always Retrieval
	Retrieval	Match	Match
		Calibration-based	
Self-RAG (7B)			
$t = 0.25$	100.0	31.9	
$t = 0.5$	23.0	10.6	31.9
$t = 0.75$	0.0	6.0	
$t =$ None	0.4	6.0	
		Model-based	
Vanilla §3			
TinyLlama (1.1B)	39.1	14.7	28.2
Phi-2 (2.7B)	**94.1**	<u>35.0</u>	36.4
Llama-2 (7B)	<u>80.3</u>	26.1	36.0
GPT-3.5	49.3	20.8	<u>38.2</u>
GPT-4*	67.6	**37.6**	**46.0**
Ours TA-ARE §4			
TinyLlama (1.1B)	54.1(+15.0)	19.0(+4.3)	28.2
Phi-2 (2.7B)	**95.5**(+1.4)	<u>36.0</u>(+1.0)	36.4
Llama-2 (7B)	86.0(+5.7)	30.7(+4.6)	36.0
GPT-3.5	<u>86.3</u>(+37.0)	35.8(+15.0)	<u>38.2</u>
GPT-4*	83.2(+15.6)	**46.4**(+8.8)	**46.0**
Average gain	+14.9	+6.7	–

Table 1: Retrieval and match accuracy on `RetrievalQA`. * indicates using 250 examples for testing to reduce API costs. Best scores in **Bold** and second best in <u>underline</u>.

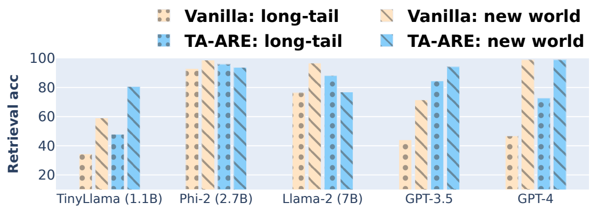

Figure 2: Retrieval accuracy between *long-tail* vs. *new world* knowledge (*i.e.*, dotted vs. slash) using **Vanilla** and ours **TA-ARE** (*i.e.*, yellow vs. blue).

the off-the-shelf Contriever (Izacard et al., 2022) and author-provided top-5 documents extracted from Wikipedia where possible. For questions from ToolQA, we use the author-provided vector database for retrieval of synthesized agendas. Otherwise, we use top-5 documents returned by Google search for new world knowledge questions. To reduce API costs, for GPT-4, we randomly select 50 data instances from each source for evaluation, resulting in 250 questions. We ask LLMs to respond `"I don't know"` if they cannot answer the question. Due to page limitation, we primarily evaluate the 1,271 questions that need retrieval in the main body and provide the overall results in the Appendix A.6. More implementation details and prompt templates are in Appendix A.4 and A.7.

3.3 Results

Table 1 (top & middle) shows the retrieval accuracy and answer match accuracy for calibration-based and model-based methods. We also present the results of standard RAG, *Always Retrieval*, which can be seen as the upper bound of the baselines. We observe that: ① **RAG generally improves QA performance.** As the knowledge necessary to answer the questions is not present in LLMs, the more frequently retrieval occurs, the higher the answer accuracy becomes for all models. However, GPT-4 possesses the highest QA accuracy despite retrieving only 67.6% of the time, indicating that fully utilizing retrieved context is also crucial for generating correct answers (Asai et al., 2023a). ② **The effectiveness of Self-RAG largely depends on threshold tuning.** As shown in Table 1 (top), Self-RAG achieves high performance when setting a low retrieval threshold ($t = 0.25$) while never retrieving when the threshold is high ($t = 0.75$). This indicates that calibration-based methods require threshold tuning to find the best trade-off between task performance and retrieval efficiency. ③ **The effectiveness of vanilla prompting varies and does not scale with model sizes.** Surprisingly, Table 1 (middle) shows larger models (GPT-3.5/4) perform worse than smaller yet strong models (Phi-2/Llama-2) in retrieval accuracy, suggesting that LLMs possess a certain degree of ability to perceive their knowledge boundaries (Yin et al., 2023; Ren et al., 2023). Yet, vanilla prompting is insufficient in guiding LLMs to make reliable retrieval decisions.

3.4 Error Analysis

To investigate why vanilla prompting performs poorly for ARAG, we conduct an error analysis for GPT-3.5 and plot Fig.1. Red area indicates more than half of the time, GPT-3.5 overconfidently perceives no external information is required to answer the questions, leading to mostly incorrect predictions. Conversely, Blue area shows that, without additional information, GPT-3.5 *"knows"* it does not know the answer, therefore responding `"I don't know"`. Together, this reveals that LLMs can potentially discern the need for resource retrieval. We further find in Fig.2 that all LLMs can better recognize their lack of knowledge about the new world, leading them to actively request retrieval. However, they tend to be weak in handling long-tail questions, as depicted in Fig.2 (yellow).

4 Improving Adaptive RAG Prompting

This section presents an improved model-based ARAG method and evaluates its effectiveness.

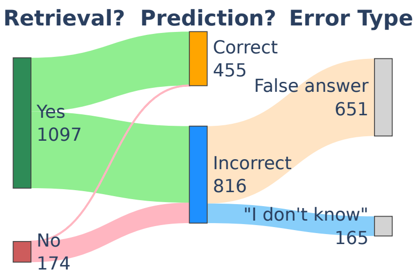

Figure 3: Error analysis of ours **TA-ART** for GPT-3.5. Compared to Fig.1, we can see that the areas of Red and Blue significantly reduce, indicating that GPT-3.5 has improved awareness of when it needs retrieval.

4.1 Method

Based on our findings in §3.4, we propose **T**ime-**A**ware **A**daptive **RE**trieval via ICL (**TA-ARE**), a simple yet effective method to improve ARAG without calibration or additional training. Given that new world knowledge questions often contain time-sensitive information (*e.g.*, "last week", "recent"), we include `"Today is current_date()"` in the instruction to enhance models' awareness of time. For long-tail knowledge, we use SimCSE (Gao et al., 2021) to select top-2 semantically closest long-tail questions answered incorrectly from the discarded set in §2, denoted as `[Yes]` demonstrations. For `[No]` demonstrations, we manually create another two questions, ensuring no extra information is required for most LLMs to answer (*e.g.*, `What is the capital of France?`).

	Time	Example	Avg. Retrieval	Avg. Match
1			65.8	24.8
2	✔		72.4	27.0
3		✔	78.9	29.3
4	✔	✔	80.6	31.1

Table 2: Ablation study for current date and demonstration examples. Results are averaged for all models.

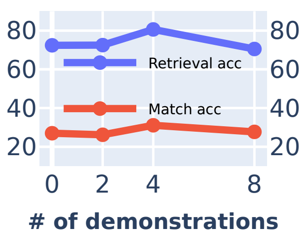

Figure 4: Effect of different numbers of demonstrations. Averaged for all models.

4.2 Results

Table 1 shows TA-ARE significantly improves all baselines, with an average gain of 14.9% and 6.7% for retrieval and QA accuracy, respectively. Fig.2 illustrates the improvement for all long-tail questions and most new world questions. As shown in Fig.3, we plot the error analysis on GPT-3.5 using our proposed TA-ARE. Compared to Fig.1 which uses vanilla prompting, we can see that the areas of Red and Blue significantly reduce, indicating that GPT-3.5 has improved awareness of when it needs retrieval, demonstrating our approach successfully elicits this ability.

In addition, our plotting enables us to conduct fine-grained error analysis for RAG. We can see that part of the LightYellow area (when `Retrieval=Yes` and `Prediction=Incorrect`) generally represents two cases: First, the retrieved documents are noisy and might not contain relevant information to answer the questions. Thus, LLMs cannot make correct predictions; Second, the retrieved documents contain necessary information, but LLMs cannot fully utilize them and make correct predictions. While this is out of the scope of this paper, future works are required to make RAG systems more robust and effective (Asai et al., 2023a; Yoran et al., 2023).

4.3 Ablation Study

The ablation studies in Table 2 validate the effectiveness of TA-ARE: time awareness and relevant in-context demonstrations help LLMs decide the necessity of retrieval for new world and long-tail questions. Table 8 and Table 9 in the Appendix further show the fine-grained performance of each model. We further evaluate the number of in-context demonstrations in Fig.4, showing that 4 demonstrations, comprising 2 `[Yes]` and 2 `[No]` examples, have the best performance.

5 Conclusion

This paper presents a new dataset `RetrievalQA` to assess adaptive RAG for short-form open-domain QA. We find vanilla prompting is insufficient in guiding LLMs to make reliable retrieval decisions. As an initial attempt, we propose TA-ARE, a simple yet effective method to help LLMs assess the necessity of retrieval, obviating the need for calibration or additional training.

Limitations

We identify the limitations of our work as follows:

- We mainly collect data from existing data sources and use GPT-4 for filtering out answerable questions. While we have done preliminary human checking in Appendix A.5, it is possible that some questions in the dataset do not require additional information for LLMs to answer. Future work could develop advanced algorithms to do more efficient and rigour filtering.
- We primarily focus on short-form QA in this paper and do not assess long-form generation tasks. It should be noted that methods, including Jiang et al. (2023); Asai et al. (2023b), are capable of long-form generation tasks. Self-RAG can also perform sophisticated self-reflection, which goes beyond adaptive retrieval.
- We acknowledge that some of the retrieved documents may not contain the answers or the information needed to answer the questions. While improving retrieval relevance and accuracy is out of the scope of this paper, noisy context may interfere with LLMs and hurt the QA performance.
- While we find our prompt templates work well, we do not perform prompt tuning in this paper. We acknowledge that prompt templates can be sensitive to LLMs, and there are methods to find optimal prompts (Shin et al., 2020; Deng et al., 2022). We believe optimal prompts can be found and further improve performance. We leave this as a future work.

Ethical Statement

The `RetrievalQA` dataset, meticulously curated to evaluate LLMs' self-awareness ability to decide when to retrieve external resources, is constructed only using publicly available data sources. We rigorously vetted the licenses of the five publicly available datasets for compliance, ensuring that all our research methodologies aligned with institutional, national, and global ethical standards. We carefully examine the data to ensure no privacy concerns or violations. We do not collect any personally identifiable information. All data used in this paper is obtained following legal and ethical standards. In addition, we adhere to the terms of use and policies of OpenAI and Meta.

Acknowledgements

This work is supported by the Australian Research Council DP240102349. We thank anonymous reviewers for their valuable comments.

References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023a. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, Toronto, Canada. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023b. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023a. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models.

Zhangyin Feng, Weitao Ma, Weijiang Yu, Lei Huang, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting liu. 2023b. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023a. A survey for efficient open domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023b. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. ToolQA: A dataset for LLM question answering with external tools. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Appendix

A.1 Data Sources

In this work, we collect data from the following sources:

RealTimeQA (Kasai et al., 2023) A dynamic question-answering (QA) based on weekly-published news articles, which challenges static LLMs. We select data from 1 October 2023 to 12 January 2024. These data comprise weekly quizzes extracted from news websites, encompassing broad topics, including politics, business, and entertainment.

FreshQA (Vu et al., 2023) A QA benchmark with 600 questions that cover a wide range of questions and answer types. We use the fast-changing subset so that the knowledge memorized in LLMs can potentially be outdated, thus requiring external new information.

ToolQA (Zhuang et al., 2023) A benchmark to faithfully evaluate LLMs' ability to use external tools. We use questions from the Personal Agenda domain, which consists of 100 synthesized questions with virtual names and events.

PopQA (Mallen et al., 2023) An entity-centric open-domain QA dataset about entities with a wide variety of popularity. We use the long-tail subset of the data.

TriviaQA (Joshi et al., 2017) A reading comprehension dataset containing question-answer-evidence triples. We follow Asai et al. (2023b) and use the test split of the unfiltered version.

A.2 Details of `RetrievalQA` Construction

In this section, we discuss the details of our dataset construction. Our goal is to evaluate adaptive RAG (ARAG) methods and see how good they are at deciding when to retrieve. Therefore, we need the ground truth labels for each question's retrieval necessity. Ultimately, there are three kinds of questions here:

- Case 1: for all LLMs, questions that can be answered using only their parametric knowledge
- Case 2: for all LLMs, questions that can *not* be answered using only the parametric knowledge, therefore requiring external retrieval
- Case 3: questions that can be answered with their parametric knowledge for some models but can not be answered for some other models

We do not consider Case 3 because those questions cannot fairly measure whether retrieval is required for different LLMs. For Case 1, it is not trivia to collect questions that can be answered only using the parametric knowledge of LLMs. This is because different LLMs have different levels of pre-trained knowledge, and it is hard to measure (Petroni et al., 2019; Kadavath et al., 2022). For example, given a question, GPT-3.5 may fail to answer and need the help of external knowledge, while Llama-2 may answer correctly using its own knowledge because it has seen the question in the training data. The pre-training corpora are sometimes unavailable, especially for proprietary models, and we can not guarantee that the collected questions can be 100% answered with their own knowledge, as shown in the table below.

However, different from Cases 1 and 3, for Case 2, theoretically, it is possible to collect data that guarantees the knowledge to answer the questions is not present in the models. For instance, new world knowledge occurred after model training and long-tail knowledge that did not (or rarely) appear in the training corpora. Therefore, we collect 1,271 questions (Case 2) that are guaranteed cannot be answered without external information. The data collection process is detailed in §2. The dataset statistics are shown in Table 3. The examples of data instances are in Table 11.

To avoid potential bias in the evaluation towards methods that retrieve more often, we additionally collect 1,514 questions (Case 1) that can be answered using GPT-2's parametric knowledge from the discard set. Specifically, we use GPT-2 (small, 124M) in the zero-shot closed-book QA setting to evaluate the discard set. We only keep questions that can be answered using GPT-2's parametric knowledge (when the loose match score = 1), assuming that larger and stronger LLMs are also highly likely to succeed if small and weak GPT-2 can answer correctly without retrieval. We also use the entire PopQA dataset (the rest of the long-tail split), which has 12,883 data instances that are more common on the web. We found that GPT-2 cannot answer any new-world questions from the discard set, which is reasonable.

A.3 Model Details

We evaluate strong instruction-tuned models with a varying scale of model size: TinyLlama (1.1B)

Category	Data Source	# Original	# After Filtering	# Avg. Q Tokens	# Avg. Ans Tokens	# Avg. Doc Tokens (Top-5)
New world knowledge	RealTimeQA (Kasai et al., 2023)	397	188	19.0	3.1	216.7
	FreshQA (Vu et al., 2023)	127	54	13.8	3.9	227.5
Long-tail knowledge	ToolQA (Zhuang et al., 2023)	100	75	21.7	3.5	425.3
	PopQA (Mallen et al., 2023)	$1,399$	659	8.8	4.0	540.1
	TriviaQA (Joshi et al., 2017)	$7,313$	295	17.3	5.9	703.3
Total/Average	`RetrievalQA`	$9,336$	$1,271$	13.2	4.3	510.1

Table 3: Data statistics of `RetrievalQA` (questions need retrieval). **# Avg. Q, Ans, Doc Tokens** means the average number of tokens of questions, answers, and top-5 retrieved documents, respectively. We use the `tiktoken` python library to calculate the number of tokens.

(Zhang et al., 2024), Phi-2[2] (2.7B) (Gunasekar et al., 2023; Li et al., 2023), Llama-2 (7B) (Touvron et al., 2023), GPT-3.5 (OpenAI, 2022), and GPT-4 (OpenAI, 2023). We also use Self-RAG (7B, Asai et al., 2023b), which is fine-tuned based on Llama-2 7B using instruction-following corpora with interleaving passages and reflection tokens. We download the models from HuggingFace[3]. The model details, including downloading URLs, model size, and release date, can be found in Table 4.

A.4 Implementation Details

For fair comparisons, we use the same setting following Self-RAG for all experiments. The detailed hyperparameters are summarized in Table 5.

For Self-RAG, we set the retrieval threshold $t = [0.25, 0.5, 0.75, \text{None}]$. Lower thresholds encourage more frequent retrieval, while None means the model itself decides when to retrieve by generating the specific `[Retrieval]` token. Since the quality of the retrieved documents is not the focus of this paper, we use the off-the-shelf Contriever (Izacard et al., 2022) and author-provided top-5 documents extracted from Wikipedia where possible for long-tail knowledge questions. For questions from ToolQA, we use the author-provided vector database for retrieval of synthesized agendas. Otherwise, we use top-5 documents returned by Google search[4] for new world knowledge questions. To reduce API costs, for GPT-4, we randomly select 50 data instances from each source for evaluation, resulting in 250 questions. We ask LLMs to respond `"I don't know"` if they cannot answer the question.

For instruction-tuned LMs, we use the official system prompt or instruction format used during training if publicly available. We use vLLM (Kwon et al., 2023) for accelerated inference.

A.5 Sanity Check: Baselines Without Retrieval

We perform a sanity check on our `RetrievalQA` (questions need retrieval) using a simple QA template (Fig.7) without retrieval. As shown in Table 6 and Fig.1, all models achieve very poor match and F1 scores on `RetrievalQA`, indicating that it is extremely hard for models to answer the questions without consulting external resources.

We notice that TinyLlama, Phi-2, and Self-RAG have slightly better performance than larger models. Considering that these models were trained recently (as shown in Table 4), they might have learned some new knowledge and can answer some questions correctly. Additionally, we conducted human checking on the questions answered correctly and found that some questions were mismarked due to multiple possible ground truths. For example, for the question: `"Where will NeurIPS be located this year (2024)?"`, the model answers: `"NeurIPS will be held in Montreal, Canada."`, and the ground truth is an array of `["Vancouver, Canada", "Vancouver", "Canada"]`. Since the model prediction contains `Canada`, this answer was marked correct. However, **LLMs themselves still do not *truly* know the answer**. The outdated knowledge stored in their pa-

[2] We acknowledge that Phi-2 has not been instruction fine-tuned; however, we find it performs decently well in understanding instructions.

[3] `https://huggingface.co/models`

[4] We use SerpApi for Google search.

Model Name	Model Size	Release
`TinyLlama/TinyLlama-1.1B-Chat-v1.0`	1.1B	Dec 2023
`microsoft/phi-2`	2.7B	Dec 2023
`meta-llama/Llama-2-7b-chat-hf`	7B	Jul 2023
`selfrag/selfrag_llama2_7b`	7B	Oct 2023
`gpt-3.5-turbo`	–	Nov 2022
`gpt-4-turbo-preview`	–	Mar 2023

Table 4: Model used in the experiments.

Parameters	Values
`temperature`	0.0
`top_p`	1.0
`max_tokens`	100
Retrieved docs	top-5
Threshold (Self-RAG)	[None, 0.25, 0.5, 0.75]
# demonstrations (TA-ARE)	4
Eval metric	match/retrieval accuracy

Table 5: Implementation hyperparameters.

rameters makes them hallucinate. Since these questions only take a tiny portion of the entire dataset (as an example shown in Fig.1, the tiny red line from `Retrieval=No` to `Prediction=Correct`), and early-trained models such as Llama-2 and GPT-3/4 perform worse, we still keep them in our dataset.

Model	Match	F1
TinyLlama (1.1B)	4.2	1.3
Phi-2 (2.7B)	7.2	3.9
Llama-2 (7B)	2.0	0.7
Self-RAG (7B)	6.0	1.5
GPT-3.5	1.2	1.0
GPT-4*	2.4	2.3

Table 6: Match and F1 scores of models on `RetrievalQA` (1,271) **without** retrieval. * indicates that we evaluate GPT-4 using 250 examples to reduce API costs.

Model	Match
TinyLlama (1.1B)	88.1
Phi-2 (2.7B)	87.7
Llama-2 (7B)	89.8
Self-RAG (7B)	88.2
GPT-3.5	91.1
GPT-4*	88.4

Table 7: Match scores of models on 1,517 questions that do not need retrieval.

We also run the sanity check on the 1,514 questions that do not need retrieval. As shown in Table 7, even strong models like GPT-3.5 and GPT-4 can not reach 100% match accuracy using their parametric knowledge.

Baselines	Adaptive Retrieval	
	Retrieval	Match
TinyLlama (1.1B)	90.9 (+51.8)	27.5 (+12.8)
Phi-2 (2.7B)	88.7 (-5.4)	33.8 (-1.2)
Llama-2 (7B)	47.0 (-33.3)	16.7 (-9.4)
GPT-3.5	87.6 (+38.3)	36.2 (+15.4)
GPT-4	86.0 (+18.4)	46.0 (+8.4)
Average gain	+14.0	+5.2

Table 8: Ablation: our **TA-ARE** without the current date. (-red) means performance losses compared to **Vanilla** prompting in Table 1.

Baselines	Adaptive Retrieval	
	Retrieval	Match
TinyLlama (1.1B)	73.8 (+34.7)	23.1 (+8.4)
Phi-2 (2.7B)	89.5 (-4.6)	32.2 (-2.8)
Llama-2 (7B)	90.0 (+9.7)	31.4 (+5.3)
GPT-3.5	36.7 (-12.6)	18.9 (-1.9)
GPT-4	70.0 (+2.4)	39.6 (+2.0)
Average gain	+5.9	+2.2

Table 9: Ablation: our **TA-ARE** without demonstration examples.

A.6 Overall Results

In Table 1, we only evaluate 1,271 questions that need retrieval. In this section, we evaluate the entire 2,785 data, with 1,271 labelled as required retrieval and 1,514 labelled as do not require retrieval. Besides retrieval accuracy, we also report retrieval macro precision, recall, and F1. Table 10 shows the overall results. Using questions that do not need retrieval and questions that need retrieval, we have a comprehensive evaluation of ARAG methods.

A.7 Prompt Templates

We present the prompt templates used in the experiments, as shown in Fig.5, Fig.6, Fig.7, Fig.8.

Given a question, determine whether you need to retrieve external resources, such as real-time search engines, Wikipedia, or databases, to answer the question correctly. Only answer "[Yes]" or "[No]".

Question: `{question}`
Answer:

Figure 5: **Vanilla** prompt template for adaptive retrieval (§3.1).

Today is `{datetime.today()}`. Given a question, determine whether you need to retrieve external resources, such as real-time search engines, Wikipedia, or databases, to answer the question correctly. Only answer "[Yes]" or "[No]".

Here are some examples:
`{demonstration examples}`

Question: `{question}`
Answer:

Figure 6: Ours **TA-ARE** prompt template for adaptive retrieval (§4).

Please use your own knowledge to answer the questions. Only include the answer in your response and try to be concise. If you do not know the answer, just say "I don’t know".

Question: `{question}`
Answer:

Figure 7: Instruction prompt template for QA without retrieval.

Please answer the question based on the provided context. Only include the answer in your response and try to be concise. If you do not know the answer, just say "I don’t know".

Paragraph:
`{retrieved documents}`

Question: `{question}`
Answer:

Figure 8: Instruction prompt template for QA with retrieved documents.

Baselines (2,785)	No Retrieval	Adaptive Retrieval					Always Retrieval
	Match	Match	Retrieval Acc	Precision	Recall	F1	Match
			Calibration-based				
Self-RAG (7B)							
$t = 0.25$	50.7	64.3	45.6	50.0	22.8	31.3	64.3
$t = 0.5$		53.2	53.6	51.2	51.7	51.5	
$t = 0.75$		50.7	54.4	50.0	27.2	35.2	
$t =$ None		49.3	54.5	50.1	62.9	55.8	
			Model-based				
Vanilla §3							
TinyLlama (1.1B)	49.8	54.4	49.3	48.5	48.4	48.5	59.9
Phi-2 (2.7B)	51.0	64.9	48.0	51.7	55.8	53.7	65.7
Llama-2 (7B)	49.7	60.4	44.3	47.2	45.0	46.1	65.8
GPT-3.5	50.1	58.7	61.3	60.3	60.9	60.6	65.7
GPT-4*	45.4	64.4	76.0	76.0	76.8	76.4	64.2
Ours TA-ARE §4							
TinyLlama (1.1B)	49.8	56.1	44.7	45.4	45.3	45.4	59.9
Phi-2 (2.7B)	51.0	65.6	54.1	57.4	66.8	61.8	65.7
Llama-2 (7B)	49.7	63.3	44.3	47.6	44.2	45.8	65.8
GPT-3.5	50.1	65.3	67.1	68.6	70.6	69.6	65.7
GPT-4*	45.4	67.6	76.6	76.6	77.1	76.8	64.2
Average gain	–	+3.0	+1.6	+2.4	+3.4	+2.8	–

Table 10: Retrieval and match accuracy on `RetrievalQA` (overall). * indicates using 500 examples for testing to reduce API costs.

Category	Data Source	Question	Answer
New world knowledge	**RealTimeQA** (Kasai et al., 2023)	Which 2024 Republican presidential contender announced that he is ending his campaign?	Former Texas Rep. Will Hurd
	FreshQA (Vu et al., 2023)	What is the latest highest-grossing movie of the week at the Box office?	Mean Girls
Long-tail knowledge	**ToolQA** (Zhuang et al., 2023)	What time did Grace attend Broadway Show on 2022/02/17?	8:00 PM
	PopQA (Mallen et al., 2023)	What is Henry Feilden's occupation?	politician
	TriviaQA (Joshi et al., 2017)	Which bird, that breeds in northern Europe in pine and beech forests, has a chestnut brown back, grey head, dark tail, buff breast and a striped black throat?	fieldfare

Table 11: Data examples of `RetrievalQA` (questions need external retrieval).