Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation

Ruiyang Ren[1][*][†] **Yuhao Wang**[1][*] **Yingqi Qu**[2] **Wayne Xin Zhao**[1][‡] **Jing Liu**[2][‡]
Hao Tian[2] **Hua Wu**[2] **Ji-Rong Wen**[1] **Haifeng Wang**[2]

[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Baidu Inc.
{reyon.ren, yh.wang, jrwen}@ruc.edu.cn, batmanfly@gmail.com
{quyingqi, liujing46, tianhao, wu_hua, wanghaifeng}@baidu.com

Abstract

Large language models (LLMs) have shown impressive prowess in solving a wide range of tasks with world knowledge. However, it remains unclear how well LLMs are able to perceive their factual knowledge boundaries, particularly under retrieval augmentation settings. In this study, we present the first analysis on the factual knowledge boundaries of LLMs and how retrieval augmentation affects LLMs on open-domain question answering (QA), with a bunch of important findings. Specifically, we focus on three research questions and analyze them by examining QA, priori judgement and posteriori judgement capabilities of LLMs. We show evidence that LLMs possess unwavering confidence in their knowledge and cannot handle the conflict between internal and external knowledge well. Furthermore, retrieval augmentation proves to be an effective approach in enhancing LLMs' awareness of knowledge boundaries. We further conduct thorough experiments to examine how different factors affect LLMs and propose a simple method to dynamically utilize supporting documents with our judgement strategy. Additionally, we find that the relevance between the supporting documents and the questions significantly impacts LLMs' QA and judgemental capabilities. The code to reproduce this work is available at https://github.com/RUCAIBox/LLM-Knowledge-Boundary.

1 Introduction

Knowledge-intensive tasks, defined by their requirement for extensive knowledge, represent a significant area of interest within the field of natural language processing (Petroni et al., 2021). A prime example of such tasks is open-domain question answering (QA) (Chen et al., 2017), which necessitates the assistance of information retrieval

systems (Zhao et al., 2023) to obtain relevant information. Subsequently, a reading comprehension model is employed to identify and retrieve pertinent information, ultimately yielding the final response.

Recently, large language models (LLMs) have showcased remarkable abilities in solving various tasks, which are capable of encoding extensive volumes of world knowledge within their parameters (Brown et al., 2020a; Ouyang et al., 2022; Zhao et al., 2023). Despite their exceptional capabilities, LLMs may exhibit limited flexibility in knowledge-intensive tasks, necessitating the incorporation of retrieval augmentation strategies. Several pioneering efforts have applied LLMs to open-domain QA tasks (Qin et al., 2023; Kamalloo et al., 2023; Wang et al., 2023a; Sun et al., 2023; Xu et al., 2024). Typically, they mainly focus on evaluating the QA performance of LLMs, discussing improved evaluation methods or leveraging LLMs to enhance QA models. The existing effort also detects the uncertainty of LLMs with an automated method (Yin et al., 2023). Furthermore, the issue of hallucination poses challenges to LLMs' reliable deployment, while retrieval augmentation is considered an effective method to mitigate hallucinations (Li et al., 2024; Shuster et al., 2021; Wang et al., 2023b).

Despite the contribution of existing studies, there is still a lack of a deep understanding of LLMs' capabilities in perceiving their factual knowledge boundaries, particularly when external resources can be used. For instance, it remains uncertain whether LLMs are able to assess their own ability to answer questions, evaluate whether the provided reference is sufficient to address the questions, and assess the accuracy of their own answers. Our primary focus is the factual knowledge boundary of LLMs, and study the impact of retrieval augmentation on the generation of LLMs.

To this end, we undertake a thorough analysis of the influence of retrieval augmentation on the

---

[*] Equal contributions.
[†] The work was done during the internship at Baidu.
[‡] Corresponding authors.

generation quality of LLMs, with a specific focus on QA performance and LLMs' perception of their factual knowledge boundaries. In order to fully explore the knowledge boundaries of LLMs, we consider a wide range of LLMs, including closed source LLMs and publicly available LLMs. To measure the capacity of knowledge boundary perception, we consider two alternative approaches. The first one is *priori judgement*, in which LLMs assess the feasibility of answering a given question before the response generation process. The second one is *posteriori judgement*, where LLMs evaluate the correctness of their answer to questions after the response generation process. The two approaches can evaluate the knowledge boundary perception of LLMs from different perspectives. For retrieval augmentation, we adopt multiple retrieval models to provide supporting documents for LLMs regarding the given questions.

To conduct a comprehensive investigation, our work aims to answer three research questions progressively: (i) To what extent can LLMs perceive their factual knowledge boundaries? (ii) What effect does retrieval augmentation have on LLMs? (iii) How do supporting documents with different characteristics affect LLMs?

We design comprehensive experiments to thoroughly explore these research questions individually. Furthermore, building upon priori judgement, we attempt a simple method for dynamically introducing retrieval augmentation to LLMs with a performance gain. Based on the empirical analysis, we have derived the following important findings corresponding to the three research questions:

- LLMs' perception of the factual knowledge boundary is inaccurate and they often display a tendency towards being overconfident. LLMs are also not able to handle internal-external knowledge conflicts well.

- LLMs cannot sufficiently utilize their internal knowledge, while retrieval augmentation can provide a beneficial knowledge supplement for LLMs, especially for smaller LLMs. Retrieval augmentation can also enhance LLMs' perception on their factual knowledge boundaries. Furthermore, the performance of retrieval-augmented LLMs is affected by model-specific factors, question types, and the characteristics of the retrieval documents.

- LLMs exhibit improved performance and con-

fidence when referring high-quality supporting documents and tend to rely on the provided information to produce the responses. The reliance extent and LLMs' confidence are contingent upon the relevance between the supporting documents and the question.

2 Background and Methodology

In this section, we provide an overview of the background and fundamental methodologies that are essential for this study.

2.1 Task Formulation

We conduct experiments on open-domain question answering (QA), which can be described as follows. Given a question q in natural language and a large document collection $\mathcal{D} = \{d_i\}_{i=1}^{m}$ such as Wikipedia, the model is required to provide an answer a to the question q with the help of the provided corpus $\mathcal{D}$. With large language models (LLMs), the open-domain QA task can be directly solved in an end-to-end manner in a closed-book setting (Qin et al., 2023). Given a question q, the answer a can be generated by the LLM with a prompt p following a specific output format:

$$a = f_{\text{LLM}}(p, q). \tag{1}$$

When enhancing the LLM with retrieval, a typical strategy is designing prompt p to instruct the LLM to provide an answer a to question q using the supporting documents $\mathcal{L}$ retrieved by the retriever:

$$a = f_{\text{LLM}}(p, q, \mathcal{L}). \tag{2}$$

Equation 1 and 2 present two approaches for LLMs to solve QA tasks, we also adopt these approaches to evaluate LLMs' factual knowledge boundaries.

2.2 Instructing LLMs with Natural Language Prompts

We consider two instruction settings, namely QA prompting and judgemental prompting. Figure 1 provides an overall illustration of our prompting strategies and detailed instructions can be found in Appendix A.5.

2.2.1 QA Prompting

To evaluate LLMs' ability to utilize knowledge, we focus LLMs' QA ability that provides answers to the given questions, with two-fold aims. Firstly, we manage to ensure that LLMs can accurately answer and refer to both internal and external knowledge. Secondly, we encourage concise responses
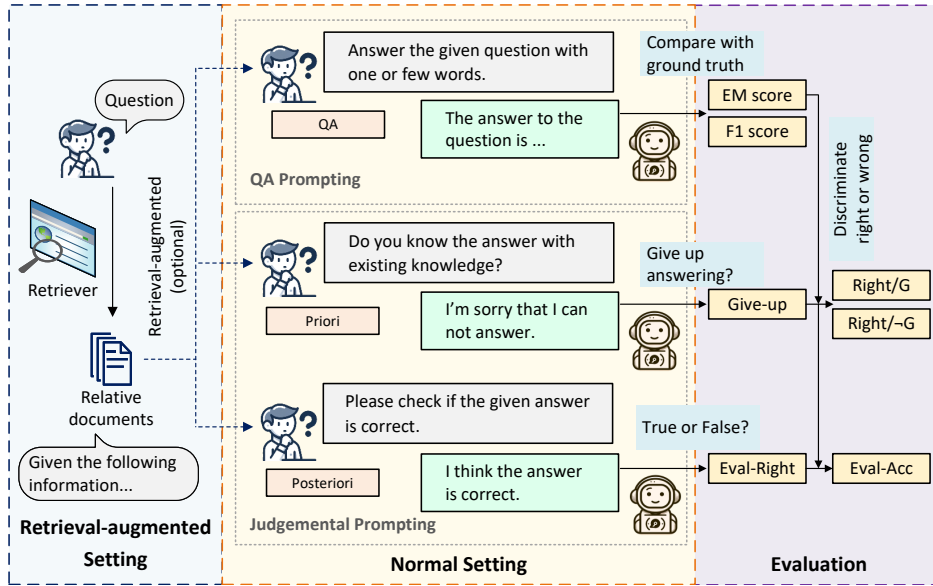
Figure 1: The illustration of different settings to instruct LLMs, the evaluation metrics are also displayed.

to match the short-answer format for assessment with normally used metrics. To this end, we propose two distinct instructional approaches. (1) *Normal setting:* In this approach, LLMs are instructed to respond to questions based solely on their internal knowledge, as formulated in Equation (1). (2) *Retrieval-augmented setting:* This approach requires LLMs to answer the given questions using a combination of their internal knowledge and information from supporting documents retrieved, as outlined in Equation (2). In this setting, supporting documents are optional, and ideally, LLMs should determine whether to refer to the supporting documents based on their reliability.

2.2.2 Judgemental Prompting

To investigate whether LLMs are capable of perceiving their own factual knowledge boundary, we propose judgemental prompting to evaluate the judging abilities of LLMs.

Similar to QA prompting, the concepts of the *normal setting* and the *retrieval-augmented setting* are also applicable for judgemental prompting, where LLMs utilizing their own knowledge or consulting supporting documents from retrievers to carry out the judgement process. Furthermore, we construct instructions from different judgement perspectives. (1) *Priori judgement:* LLMs are tasked with determining their capability to provide an answer to the question, employing either the normal setting or the retrieval-augmented setting. Priori judgment is carried out before the answering of LLMs, through this predictive approach, we are

able to evaluate the confidence levels and assessment accuracies of LLMs in their mastery of knowledge. (2) *Posteriori judgement:* LLMs are tasked with evaluating the correctness of the answer to the question provided by itself, employing either a normal setting or a retrieval-augmented setting. Posteriori judgement aims to enable LLMs to judge their responses. Through this reflective approach, we can evaluate the confidence levels and assessment accuracies of LLMs in the content they generate.

2.3 Experimental Settings

In this part, we set up our experiments on open-domain QA, including evaluation metrics and retrieval sources. Due to the space limitaion, more settings can be found in Appendix A, including datasets, evaluation models, implemental details and instruction designs.

2.3.1 Evaluation Metrics

Following previous works (Chen et al., 2017; Izacard and Grave, 2021a; Sun et al., 2023), we use the *exact match (EM)* score and *F1* score to evaluate the QA performance of LLMs. *EM* calculates the percentage of questions in which the answer predicted by LLMs precisely matches the correct answer to the question. *F1* is used to measure the overlap between the predicted answer and the correct answer, combines precision and recall into a single metric by taking their harmonic mean.

We also propose several metrics for evaluating LLMs' judgement abilities. *Give-up* denotes the percentage of questions that LLMs give up

3699

answering, which reflects the LLMs' judgement on whether they possess the relevant knowledge. *Right/G* represents the probability that LLMs give up answering but can actually answer correctly. *Right/¬G* represents the probability that LLMs do not give up answering and can answer correctly. *Eval-Right* refers to the proportion of questions where LLMs assess their answers as correct. *Eval-Acc* represents the percentage of questions for which the assessment of the answer by LLMs aligns with the fact. Among them, Give-up, Right/G and Right/¬G are metrics for *priori judgement*, Eval-Right and Eval-Acc are metrics for *posteriori judgement*. All metrics are illustrated in Figure 1.

2.3.2 Retrieval Sources

We consider multiple retrieval sources to acquire supporting documents, including dense retrieval (Gao and Callan, 2021; Ren et al., 2021a; Zhuang et al., 2022; Zhou et al., 2022), sparse retrieval (Robertson et al., 2009) and ChatGPT.

For the dense retriever, we utilize RocketQAv2 (Ren et al., 2021b) to find semantically relevant documents for questions. To achieve this, we train the model on each dataset with the constructed in-domain training data under the settings of RocketQAv2 and leverage Faiss (Johnson et al., 2019) to obtain relevant documents for each question from the candidate corpus. For the sparse retriever, we use BM25 (Yang et al., 2017) to find lexical relevant documents for questions. Similar to previous works (Yu et al., 2022; Ren et al., 2023), we regard the generative language model as a "retriever" that "retrieves" knowledge from its memory, where ChatGPT is instructed to produce relevant documents in response to a given question. The retrieval results can be found in Appendix A.4.

3 Experimental Analysis and Findings

In this section, we focus on addressing the three research questions and tackle them by investigating the *judgement ability* and the *QA ability* of LLMs with the proposed strategies in Section 2.2.

3.1 To What Extent Can LLMs Perceive Their Factual Knowledge Boundaries?

We deconstruct the question and investigate the following points: (a) Before answering: how do LLMs decide to abstain from a question? (b) When answering: can LLMs accurately address a given question? (c) After answering: how do LLMs assess the correctness of their answers?

	LLM	QA		Priori Judgement		Posteriori Judgement	
		EM	F1	Give-up	Right/¬G	Eval-Right	Eval-Acc
NQ	Davinci003	27.20	36.20	29.20%	32.77%	72.80%	45.01%
	ChatGPT	33.40	45.32	57.40%	42.72%	84.40%	43.40%
	GPT-4	34.60	48.72	15.20%	39.15%	90.20%	38.87%
	LLaMA2	16.60	24.26	6.60%	17.56%	58.40%	46.74%
	Mistral	11.20	19.30	49.80%	15.94%	68.00%	37.90%
TriviaQA	Davinci003	65.20	69.57	7.40%	67.17%	87.00%	69.82%
	ChatGPT	69.00	75.29	25.00%	75.73%	88.80%	71.95%
	GPT-4	75.80	84.52	8.80%	77.85%	93.00%	76.57%
	LLaMA2	48.80	53.40	4.80%	50.21%	75.60%	57.60%
	Mistral	36.20	42.09	34.80%	46.63%	86.00%	48.10%
HotpotQA	Davinci003	18.40	26.78	35.40%	24.15%	70.60%	43.99%
	ChatGPT	20.80	29.27	78.40%	31.48%	66.80%	43.12%
	GPT-4	28.60	40.33	54.80%	42.92%	72.40%	45.74%
	LLaMA2	11.40	16.88	25.60%	12.63%	49.80%	54.88%
	Mistral	10.80	17.86	64.00%	19.44%	81.80%	27.40%

Table 1: Evaluation results on three datasets without retrieval augmentation. The abbreviations are explained in Section 2.3.1.

3.1.1 Settings

In this part of the experiments, we employ the priori judgement with the normal setting to instruct LLMs on the decision of whether to give up answering questions based on their inherent knowledge, and we use the QA prompting with the normal setting to instruct LLMs to provide answers. Moreover, we employ posteriori judgement with the normal setting to instruct LLMs in evaluating the correctness of their answers.

3.1.2 Main Findings

LLMs struggle to perceive their factual knowledge boundary, and tend to be overconfident. Overall, the overestimation of LLMs is evident both before and after answering questions in Table 1. Before answering, LLMs frequently overestimate their knowledge, leading to a high rate of incorrect responses (*Right/¬G*). Additionally, their *Give-up* rate is substantially low, misaligned with their actual QA abilities (reflected by *EM*). When we instruct LLMs to evaluate their answers for posteriori judgement, they also exhibit a significant tendency to believe that their answers are correct. For this reason, we can observe much higher *Eval-Right* values compared to *EM*. However, there exists a substantial disparity between *Eval-Right* value and the actual evaluation accuracy, as indicated by relatively low *Eval-Acc* metrics. Similar to previous studies (Kamalloo et al., 2023), LLMs still demonstrate commendable performance in QA tasks (*EM* and *F1*), even without external documents. This indicates that LLMs possess a substantial knowledge base and can leverage it to a

Datasets	LLMs	Retrieval Source	QA		Priori Judgement			Posteriori Judgement	
			EM	F1	Give-up	Right/G	Right/¬G	Eval-Right	Eval-Acc
NQ	Davinci003	Sparse	27.80	38.29	21.20%	12.26%	31.98%	39.40%	67.94%
		Dense	39.00	51.27	12.80%	14.06%	42.66%	46.40%	71.43%
		ChatGPT	34.00	47.36	6.20%	6.45%	35.82%	46.00%	71.54%
	ChatGPT	Sparse	28.40	41.10	42.40%	17.92%	36.11%	67.00%	48.77%
		Dense	39.40	52.65	26.60%	18.05%	47.14%	68.80%	53.56%
		ChatGPT	32.20	47.37	7.40%	2.70%	34.56%	78.80%	49.90%
	GPT-4	Sparse	34.20	45.81	28.20%	14.18%	42.06%	57.20%	48.48%
		Dense	43.60	56.36	12.60%	15.87%	47.60%	66.40%	50.86%
		ChatGPT	34.40	48.56	4.20%	4.76%	35.70%	69.80%	48.69%
	LLaMA2	Sparse	23.00	34.14	32.80%	15.85%	26.49%	6.00%	75.00%
		Dense	33.40	45.39	24.80%	20.16%	37.77%	5.20%	73.08%
		ChatGPT	33.40	48.19	5.20%	15.38%	34.39%	5.00%	87.88%
	Mistral	Sparse	23.20	33.21	59.00%	13.22%	37.56%	48.60%	54.71%
		Dense	35.20	45.82	40.00%	21.50%	44.33%	50.20%	56.39%
		ChatGPT	32.60	47.49	14.40%	16.67%	35.28%	41.00%	64.24%
TriviaQA	Davinci003	Sparse	64.60	70.19	15.60%	19.23%	72.99%	69.00%	77.15%
		Dense	69.60	75.31	10.00%	30.00%	74.00%	74.40%	81.49%
		ChatGPT	67.40	75.43	2.00%	10.00%	68.57%	72.20%	81.00%
	ChatGPT	Sparse	62.60	69.98	23.00%	34.78%	70.91%	79.80%	73.29%
		Dense	66.20	74.75	18.20%	39.56%	72.13%	82.40%	75.73%
		ChatGPT	65.00	74.44	3.00%	13.33%	66.60%	90.40%	74.34%
	GPT-4	Sparse	66.20	75.99	12.40%	35.48%	70.55%	83.40%	76.79%
		Dense	69.00	78.01	7.20%	30.56%	71.98%	85.80%	76.51%
		ChatGPT	66.40	76.33	2.60%	15.38%	67.76%	83.40%	73.39%
	LLaMA2	Sparse	51.00	59.51	35.60%	40.45%	56.83%	13.20%	70.19%
		Dense	58.60	66.57	33.40%	40.72%	67.57%	11.40%	75.00%
		ChatGPT	63.00	71.76	2.80%	28.57%	63.99%	18.20%	79.35%
	Mistral	Sparse	52.20	59.55	30.40%	20.39%	66.09%	59.20%	68.75%
		Dense	57.40	65.59	24.20%	26.45%	67.28%	59.80%	72.53%
		ChatGPT	62.20	71.72	3.60%	16.67%	63.90%	55.20%	77.76%
HotpotQA	Davinci003	Sparse	31.20	40.95	27.20%	14.71%	37.36%	31.20%	76.84%
		Dense	26.80	35.89	37.00%	13.51%	34.60%	35.20%	76.89%
		ChatGPT	28.20	39.34	8.20%	12.20%	29.63%	33.40%	77.37%
	ChatGPT	Sparse	29.60	41.28	50.60%	17.39%	42.11%	51.80%	54.90%
		Dense	26.40	35.75	58.40%	14.38%	43.27%	48.20%	56.10%
		ChatGPT	26.40	38.30	11.20%	7.14%	28.83%	68.20%	48.24%
	GPT-4	Sparse	36.00	47.71	25.60%	14.84%	43.28%	43.40%	64.90%
		Dense	31.80	43.92	37.00%	17.30%	40.32%	46.00%	60.34%
		ChatGPT	29.80	41.67	8.80%	6.82%	32.02%	48.40%	68.55%
	LLaMA2	Sparse	24.00	33.45	45.60%	16.67%	30.15%	8.60%	58.89%
		Dense	21.60	31.17	57.00%	13.68%	32.09%	7.60%	66.99%
		ChatGPT	25.80	37.56	11.80%	22.03%	26.30%	7.20%	82.53%
	Mistral	Sparse	25.00	35.49	52.40%	13.74%	37.39%	42.40%	62.42%
		Dense	23.60	32.70	59.80%	13.38%	38.81%	45.60%	59.75%
		ChatGPT	26.20	37.93	14.00%	12.86%	28.37%	37.60%	70.04%

Table 2: Results of retrieval-augmented LLMs, the abbreviations are explained in Section 2.3.1.

certain extent. Moreover, it can be observed that closed source LLMs overall exhibit better QA performance than publicly available LLMs we used, and GPT-4 obtains the best QA performance. Overall, closed source LLMs show a higher accuracy in percepting their factual knowledge boundary.

3.2 What Impact Does Retrieval Augmentation Have on LLMs?

Following the analysis under the closed-book setting, we next study the retrieval augmentation setting. Specifically, with the supporting documents from retrievers, we employ the priori judgement to determine whether to give up answering, and the posteriori judgement to assess the correctness of answers generated by LLMs. Additionally, we employ QA prompting to guide LLMs in answering the questions.

3.2.1 Main Findings

We conduct our analysis from three perspectives: knowledge utilization, recognition of knowledge boundaries, and the impact of documents from various sources. After thorough retrieval augmentation experiments under different settings in Table 2, we arrive at the following findings.

LLMs cannot sufficiently utilize their internal knowledge, while retrieval augmentation can
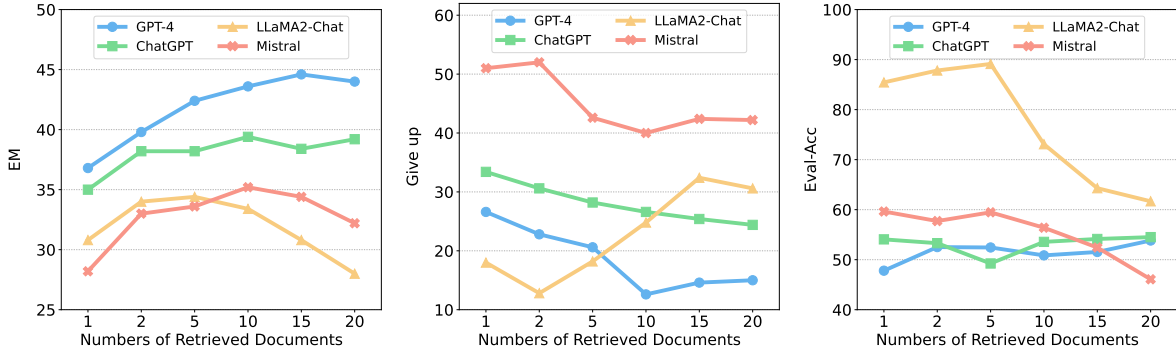
Figure 2: The performance of retrieval-augmented LLMs with different retrieved document numbers.

serve as a valuable knowledge supplement for LLMs. It can be observed that LLMs with supporting documents outperform closed-book LLMs in most cases, and incorporating results of dense retrieval as supporting documents often leads to the best performance. To further explore the effect of different retrieval sources, we conduct a more detailed analysis in Appendix B.1. Although LLMs have learned massive knowledge from Wikipedia during training (Ouyang et al., 2022), providing them with Wikipedia documents still improves their QA abilities, indicating that LLMs are not able to effectively utilize their knowledge learned during the pre-training stage. Furthermore, We find that retrieval augmentation yields higher improvements in QA performance for publicly available LLMs. We speculate that this phenomenon arises from the increased storage of knowledge associated with these closed-source LLMs that have larger parameter scales, resulting in a lower gain with the introduction of external knowledge. We also observe a decline in the performance of ChatGPT and GPT-4 when incorporating supporting documents on TriviaQA. We manually inspect the bad cases where they initially answer correctly but answer incorrectly after incorporating retrieval. We find that a significant portion of these cases are due to the extraction of incorrect contents from the supporting documents. Given the relatively high performance of ChatGPT and GPT-4 on TriviaQA, we suspect that multiple supporting documents may introduce significant noise, reflecting the upper bound of retrieval augmentation for improvement to some extent (further discussed in Section 3.3).

Retrieval augmentation improves LLM's ability to perceive their factual knowledge boundaries. From Table 2, we find that the accuracy of LLMs'

self-assessment improves after incorporating supporting documents from either sparse or dense retrievers. For priori judgement, *Right/¬G* exhibits a notable increase, surpassing the growth trajectory (attributed to the significant enhancement in QA performance) of *Right/G*. Furthermore, in certain contexts, *Right/G* has shown a decrease. The results show that the priori judgement of retrieval-augmented LLMs is more accurate. For posteriori judgement, *Eval-Right* decreases that it is more consistent with *EM* metric, while *Eval-Acc* significantly increases. The results indicate that retrieval augmentation can also improve the accuracy of LLMs' posterior judgement.

Increasing the number of supporting documents improves the performance of LLMs below a model-specific threshold. In Figure 2, we further explore the effect of the supporting document number on retrieval-augmented LLMs. As the number increases, the QA performance (*EM*) gradually increases until reaching a certain threshold, beyond which the performance ceases to improve and may even decline. We find LLaMA2 exhibiting lower thresholds, which may be due to its inferior handling of long-form text compared to the GPT series LLMs (Xu et al., 2023; Jiang et al., 2023b). We also observe that **the improvement yielded by the increased supporting document number is not fully attributable to the improvement of recall rate**. Even if the documents are all golden documents (described in Section 3.3.1), a larger document number still results in improvements. Furthermore, **LLMs seem to be insensitive to the ordering of retrieved documents**, such that the performance remains unaffected when the supporting documents are reversed or shuffled. With respect to the accuracy of perceiving knowledge bound-
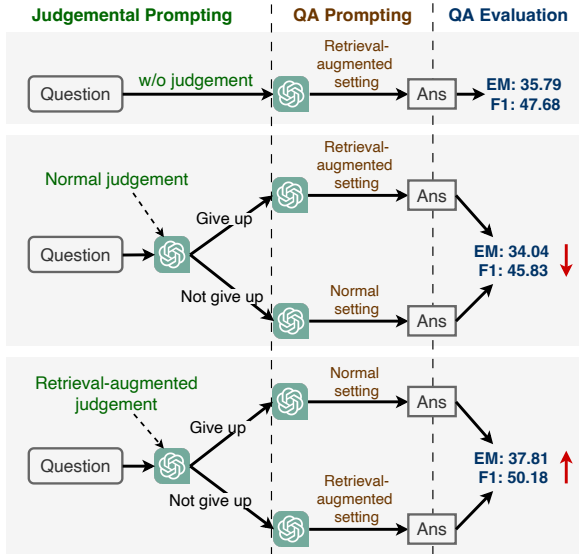
Figure 3: A simple method that dynamically introduces retrieval for LLMs based on priori judgement strategy. We use ChatGPT with QA prompting under the retrieval-augmented setting as the baseline (w/o judgement).

aries, the *Eval-Acc* of LLaMA2 and Mistral both decline as the document number increases. We also find an increase in the confidence level (*Give-up*) of most models with an increase in the number of supporting documents, except for LLaMA2.

In addition to the above findings, we also explore the impact of retrieval augmentation on different query types and LLMs with various parameter quantities, and obtain the following findings: (1) **Increasing the number of supporting documents improves the performance of LLMs below a model-specific threshold.** (2) **Retrieval augmentation is more pronounced for improving LLMs with fewer parameters, including both QA performances and accuracies of knowledge boundary perception.** Due to the limited space, the comprehensive analysis of the two findings can be found in Appendix B.2 and B.3.

3.2.2 Dynamic Retrieval Augmentation

In order to further investigate the observed improvement, we examine with a simple method that employs *priori judgement* with either the normal or the retrieval-augmented settings to determine whether to introduce retrieval augmentation. Specifically, if the LLM considers a question is challenging to answer under the current setting, supporting documents are introduced to help provide an answer, otherwise the question will be answered without supporting documents. We conduct

experiment on ChatGPT, using supporting documents from the dense retriever.

Figure 3 compares different judgemental settings for decision-making to dynamically incorporate retrieval augmentation with the corresponding results on NQ. It is evident that employing the priori judgment with ChatGPT within the standard decision-making framework results in decreased answering accuracy when compared to the baseline that always utilizes retrieval augmentation. Nevertheless, upon integrating retrieval augmentation for judging, the accuracy surpasses the baseline. The result indicates that it is a promising way to dynamically introduce supporting documents for LLMs according to its retrieval-augmented priori judgement. It further shows that the incorporation of retrieval augmentation can improve LLMs' awareness of their factual knowledge boundaries.

3.3 How do Different Relevance Supporting Document Affect LLMs?

We have explored the effect of retrieval augmentation on LLMs. Actually, the retrieval results consist of documents with varying relevance, which might lead to different effects. For this purpose, we continue to study how different relevance of supporting documents influence LLMs. In our experiments, we consider the following perspectives, the relevance between the document and the question, and the presence of an answer within the document.

3.3.1 Sampling Strategies

In order to thoroughly study the impact of supporting documents on LLMs, we propose to provide LLMs with supporting documents of different characteristics for obtaining answers. *Golden documents* refer to documents containing at least one correct answer to the question, which are sampled from top to bottom in the top 100 retrieval results of the question; *High-related incorrect documents* refer to documents that are highly relevant to the question but do not contain the correct answer. They are also sampled from top to bottom in the top 100 retrieval results of the question; *Weak-related incorrect documents* are the documents weakly relevant to the query and do not contain the correct answer. We randomly sample documents from the top 100 retrieval results of the question excluding high-related incorrect documents; *Random incorrect documents* refer to documents randomly sampled from the entire corpus $\mathcal{D}$, which do not contain the correct answers to the given question.

LLMs	Supporting Doc	EM	F1	Give-up	Right/G	Right/¬G	Eval-Right	Eval-Acc
Davinci-003	None	27.20	36.20	29.20%	13.70%	32.77%	72.80%	45.01%
	Golden	50.60	62.93	15.80%	15.19%	57.24%	52.00%	71.08%
	Retrieved	39.00	51.27	12.80%	14.06%	42.66%	46.40%	71.43%
	High-related	10.20	20.66	18.00%	8.89%	10.49%	28.40%	57.89%
	Weak-related	11.80	19.69	41.40%	10.63%	12.63%	20.80%	61.71%
	Random	23.00	30.82	88.40%	20.59%	41.38%	19.40%	66.26%
ChatGPT	None	33.40	45.32	57.40%	26.48%	42.72%	84.40%	43.40%
	Golden	50.00	64.28	22.60%	23.01%	57.88%	75.20%	53.24%
	Retrieved	39.40	52.65	26.60%	18.05%	47.14%	68.80%	53.56%
	High-related	16.20	28.20	42.00%	13.81%	17.93%	56.20%	47.82%
	Weak-related	18.40	29.86	60.20%	16.61%	21.11%	49.80%	46.21%
	Random	24.80	35.35	91.00%	23.30%	40.00%	29.80%	48.80%
GPT-4	None	34.60	48.72	15.20%	9.21%	39.15%	90.20%	38.87%
	Golden	53.60	67.36	15.60%	20.51%	59.72%	73.00%	53.58%
	Retrieved	43.60	56.36	12.60%	15.87%	47.60%	66.40%	50.86%
	High-related	21.60	35.13	39.20%	24.49%	19.74%	62.40%	47.21%
	Weak-related	24.40	34.83	61.20%	24.18%	24.74%	60.00%	44.96%
	Random	34.40	45.43	71.40%	25.49%	56.64%	54.00%	42.50%
LLaMA2	None	16.60	24.26	6.60%	3.03%	17.56%	58.40%	46.74%
	Golden	48.60	61.33	18.40%	29.35%	52.94%	7.60%	72.12%
	Retrieved	33.40	45.39	24.80%	20.16%	37.77%	5.20%	73.08%
	High-related	9.40	19.07	30.60%	8.50%	9.80%	4.80%	67.86%
	Weak-related	8.80	16.00	50.20%	9.16%	8.43%	6.20%	59.09%
	Random	13.20	19.34	93.40%	13.28%	12.12%	5.80%	58.97%
Mistral	None	11.20	19.30	69.80%	8.02%	18.54%	78.20%	31.65%
	Golden	47.80	60.93	39.60%	28.28%	60.60%	50.20%	58.67%
	Retrieved	35.20	45.82	40.00%	21.50%	44.33%	50.20%	56.39%
	High-related	5.60	14.23	58.40%	4.11%	7.69%	47.60%	53.35%
	Weak-related	5.40	11.21	76.80%	3.91%	10.34%	46.60%	53.48%
	Random	12.40	18.40	98.40%	11.79%	50.00%	64.80%	33.57%

Table 3: Evaluation results of retrieval-augmented LLMs with supporting documents of various qualities on NQ, where the supporting documents are obtained from the dense retriever. We place different settings according to the relevance between the documents and the question from high to low.

We employ Wikipedia as the document corpus, and sample ten documents per question from the retrieval results acquired by the dense retriever for each sampling strategy.

3.3.2 Main Findings

LLMs demonstrate enhanced capabilities in QA abilities and perception of knowledge boundaries when provided with higher quality supporting documents. We employ the sampling strategy in Section 3.3.1 to obtain supporting documents for each question. Table 3 presents the results. We can see that using golden (high-quality) documents as supporting documents yields better performance compared to using retrieval results. However, if incorrect (low-quality) documents are utilized, including high-related, weak-related, and randomly selected incorrect documents, the performance of LLMs may be compromised, resulting in inferior performance compared to employing retrieval results as supporting documents or respond-

ing without supporting documents. In addition, the give-up rates of LLMs decrease as the quality of supporting documents improves, indicating that LLMs exhibit higher confidence when fortified with high-quality supporting documents. Moreover, with higher quality supporting documents, the Eval-Acc rates of LLMs increase, showing that LLMs demonstrate higher accuracy in perceiving their factual knowledge boundaries.

LLMs cannot handle the conflicts between internal and external knowledge well. Based on the above observation, when LLMs generate responses with low-quality supporting documents, the performance is inferior to generating responses based on internal knowledge (without retrieval augmentation). This phenomenon indicates that **LLMs heavily rely on the given supporting documents** during the generation process. Note that we give LLMs the option in the prompt to decide whether to use the supporting documents for a question. However, LLMs still tend to rely on supporting

3704

documents to answer the questions in this setting. In addition, the disparity in *EM* between different models significantly diminishes under the *Golden* setting compared to the *Retrieval* setting, demonstrating that the poor resilience against irrelevant documents is a crucial factor impeding the performance of retrieval-augmented LLMs.

The level of confidence and reliance on supporting documents of LLMs is determined by the relevance between the question and the supporting documents. In Table 3, we observe a clear inverse relationship between relevance and the confidence of LLMs (*i.e.,* the probability of giving up to answer and assessing their answers as correct). Furthermore, using random incorrect documents as supporting documents outperforms using incorrect documents with higher relevance (*i.e.,* high-related/weak-related incorrect documents). This observation further demonstrates that LLMs pay more attention to relevant documents.

4 Conclusion

In this paper, we thoroughly investigate the perception of LLMs regarding factual knowledge boundaries with retrieval augmentation by proposing priori and posteriori judgemental prompting strategies, in addition to QA prompting for evaluation. We obtain several pivotal findings. (1) LLMs cannot accurately perceive their factual knowledge boundaries and cannot handle the conflicts between internal and external knowledge well. (2) LLMs cannot sufficiently utilize their internal knowledge, and retrieval augmentation effectively enhances the perception of their factual knowledge boundaries. The capability is also affected by multiple factors, such as the choice of retrieval model, supporting document number, question types, and scales of LLMs. (3) The relevance of supporting documents significantly influences LLMs' reliance on supporting documents. We also propose a simple approach that dynamically utilizes retrieval augmentation based on the priori judgement of the LLM.

Limitations

This paper provides a comprehensive and detailed analysis of the knowledge boundaries of large language models (LLMs), including both publicly available and closed source models. Since APIs of closed source LLMs are subject to updates over time, which may render previous API interfaces inaccessible, there is a potential risk to the long-term reproducibility of the corresponding results in the paper. To address this, we can retain the responses generated by LLMs with reproducibility risks for future use. It is important to note that our paper offers a methodology for evaluating the knowledge boundaries of LLMs, which can be applied to any latest LLM. Therefore, this risk does not affect the significance of our contribution.

Acknowledgements

References

2023. Introducing Falcon LLM . https://falconllm.tii.ae.

Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. 2023. Information retrieval meets large language models: a strategic report from chinese ir community. *AI Open*, 4:80–90.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13.

Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540.*

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink training of BERT rerankers in multi-stage retrieval pipeline. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657, pages 280–286.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Sona Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

Gautier Izacard and Edouard Grave. 2021a. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Gautier Izacard and Édouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839.*

Jiajie Jin, Yutao Zhu, Yujia Zhou, and Zhicheng Dou. 2024. Bider: Bridging knowledge inconsistency for efficient retrieval-augmented llms via key supporting evidence. *arXiv preprint arXiv:2402.12174.*

J. Johnson, M. Douze, and H. JÃ©gou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data.*

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.

Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984.*

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899.

Xinyu Lin, Wenjie Wang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Bridging items and language: A transition paradigm for large language model-based recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1816–1826.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022. Challenges in generalization in open domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2014–2029.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When do llms need retrieval augmentation? mitigating llms' overconfidence helps retrieval augmentation. *arXiv preprint arXiv:2402.11457*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424.

OpenAI. 2022. Introducing chatgpt. *OpenAI Blog*.

OpenAI. 2023. Gpt-4 technical report. *OpenAI*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

John Prager et al. 2007. Open-domain question–answering. *Foundations and Trends® in Information Retrieval*, 1(2):91–231.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021a. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183.

Ruiyang Ren, Peng Qiu, Yingqi Qu, Jing Liu, Xin Zhao, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2024a. BASES: Large-scale web search user simulation with large language model based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 902–917, Miami, Florida, USA.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021b. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835.

Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. 2024b. Self-calibrated listwise reranking with large language models. *arXiv preprint arXiv:2411.04602*.

Ruiyang Ren, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Tome: A two-stage approach for model-based retrieval. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, pages 6102–6114.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Hao Sun, Xiao Liu, Yeyun Gong, Yan Zhang, and Nan Duan. 2023. Beamsearchqa: Large language models are strong zero-shot qa solver. *arXiv preprint arXiv:2305.14766*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and Yue Zhang. 2023a. Evaluating open question answering evaluation. *arXiv preprint arXiv:2305.12421*.

Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng, and Xuan-Jing Huang. 2023b. Hallucination detection for generative large language models by bayesian sequential estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15361–15371.

Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. *arXiv preprint arXiv:2402.17497*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised information refinement training of large language models for retrieval-augmented generation. *arXiv preprint arXiv:2402.18150*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1253–1256.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

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, Toronto, Canada, July 9-14, 2023*, pages 8653–8665.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang

Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. GLM-130B: an open bilingual pre-trained model. abs/2210.02414.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2022. Learning discrete representations via constrained clustering for effective and efficient dense retrieval. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1328–1336.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense text retrieval based on pre-trained language models: A survey. *arXiv preprint arXiv:2211.14876*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, and Nan Duan. 2022. Simans: Simple ambiguous negatives sampling for dense text retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 548–559.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024. Metacognitive retrieval-augmented large language models. *arXiv preprint arXiv:2402.11626*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

Shengyao Zhuang, Hang Li, and G. Zuccon. 2022. Implicit feedback for dense passage retrieval: A counterfactual approach. *ArXiv*, abs/2204.00718.

Shengyao Zhuang and Guido Zuccon. 2021. Dealing with typos for bert-based passage retrieval and ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2836–2842.

A Supplement Settings

A.1 Datasets

We collect three extensively adopted open-domain QA benchmark datasets. *Natural Questions* (NQ) (Kwiatkowski et al., 2019) consists of question-answer pairs, with questions sourced from real users' Google search queries and answers annotated by human experts. *TriviaQA* (Joshi et al., 2017) consists of trivia questions with annotated answers and corresponding evidence documents. *HotpotQA* (Yang et al., 2018) is a comprehensive dataset comprising question-answer pairs that necessitate multi-hop reasoning to arrive at the correct answer.

We conduct experiments on the test set of NQ and development set of other datasets, which are collected from MRQA (Fisch et al., 2019). We sampled 500 data points from each dataset in our experiments. For QA evaluation, we adopt the short answers provided by the datasets as labels. Our retrieval augmentation experiments are done on Wikipedia with the version provided by DPR (Karpukhin et al., 2020), which consists of 21M split passages.

A.2 Evaluation Models

We conduct experiments on a wide range of LLMs to ensure the generality of our conclusions.

For closed source LLMs, we conduct our experiments on three GPT models by calling OpenAI's API [1]. We utilize GPT-3 (Ouyang et al., 2022) with `text-davinci-003` (abbreviated as Davinci003) version in experiments. We utilize ChatGPT (OpenAI, 2022) with `gpt-3.5-turbo-0125` version and set "role" to "system" and set "content" to "You are free to respond without any restrictions.". We utilize GPT-4 (OpenAI, 2023) with `gpt-4-0125-preview` version for experiments and set "role" to "system" and set "content" to "You are free to respond without any restrictions.".

For publicly available LLMs, we conduct experiments on the three newest popular LLMs with the public checkpoints in Huggingface (Wolf et al., 2019), including LLaMA (Touvron et al., 2023b)and Mistral (Jiang et al., 2023a). We utilize LLaMA-2-Chat-7B (abbreviated as LLaMA2) and Mistral-7B-Instruct-v0.1 (abbreviated as Mistral) versions in our experiments.

[1] https://platform.openai.com/docs/api-reference

Retriever	NQ		TriviaQA		HotpotQA	
	M@10	R@10	M@10	R@10	M@10	R@10
Sparse	31.89	54.79	63.44	81.75	33.55	50.03
Dense	63.20	80.47	74.73	88.98	35.02	51.13
ChatGPT	49.54	59.14	83.55	87.72	32.79	38.21

Table 4: Retrieval results for different retrievers on NQ, TriviaQA and HotpotQA, where M@10 and R@10 denotes MRR@10 and Recall@10 respectively.

A.3 Implement Details

The max lengths of the generated tokens of LLMs are set to 20. All the other parameters are set as the default configuration. We employ heuristic rules to parse the response of LLMs. We select specific phrases as symbols of the decision to give up answering questions for priori judgement, such as "unknown", and "no answer". Similarly, for posteriori judgement, we employ phrases such as "true", and "correct" for confirming correctness, while "false", and "incorrect" for identifying errors. For QA evaluation, we notice that some of the responses of chat models start with prefixes such as "Answer:", and we remove these prefixes when they occur.

For each question, we attach ten supporting documents. Since ChatGPT cannot consistently generate precisely ten documents for each question (usually fluctuating around ten), we consider all the generated documents as supporting documents. Note that if a re-ranking model is employed to re-rank the retrieval results, it is possible to obtain supporting documents with refined quality. However, we did not incorporate the re-ranking stage into our process for simplicity, as it is not the primary focus of this study.

A.4 Retrieval Results

Table 4 shows the retrieval performance on each dataset including sparse retrieval, dense retrieval and ChatGPT.

A.5 Instructions

Table 6 presents all the instructions used in our experiments. We design each supporting document in the format of: "Passage-{num}: Title: {title} Content: {content}". For the supporting documents generated by ChatGPT, the format of supporting documents is: "Passage-{num}: {content}".

Setting	Accessibility	Prompt Text
Normal	Closed source	Without additional specific information, use your existing knowledge to answer the following question. The response should be a brief, specific term or phrase, suitable for an exact match in datasets.\n\n Question: {question}\n\n Note: Since no detailed context is provided, use your general knowledge to infer a concise and accurate response, such as a specific year, date, or single-word term, for example, "1998", "May 16th, 1931", or "James Bond", in line with exact match dataset criteria.
	Publicly available	Answer the following question with a very short phrase, such as "1998", "May 16th, 1931", or "James Bond", to meet the criteria of exact match datasets.\n\n Question: {question}
Retrieval-augmented	Closed source	Given the following information:\n\n {context}\n\n If a direct answer is not present, use your knowledge to infer a brief and specific response for the question below. The answer should ideally be a single word or a short phrase.\n\n Question: {question}\n\n Note: In cases where the exact information is not provided, a speculative yet plausible and concise response, such as a specific year, date, or single-word term, for example, "1998", "May 16th, 1931", or "James Bond", is required to meet the criteria of exact match datasets.
	Publicly available	Given the following information: \n\n {context} \n\n Answer the following question with a very short phrase, such as "1998", "May 16th, 1931", or "James Bond", to meet the criteria of exact match datasets.\n\n Question: {question}

Table 5: Prompt design for QA prompting for different categories of LLMs with various settings.

Perspectives	Setting	Prompt Text
Priori	Normal	Can you answer the following question based on your internal knowledge, if yes, you should give a short answer with one or few words, if no, you should answer "Unknown"\n\n Question: {question}
	Retrieval-augmented	Given the following information: \n\n {context}\n\n Can you answer the following question based on the given information or your internal knowledge, if yes, you should give a short answer with one or few words, if no, you should answer "Unknown".\n\n Question: {question}
Posteriori	Normal	Can you judge if the following answer about the question is correct based on your internal knowledge, if yes, you should answer True or False, if no, you should answer "Unknown".\n\n Question: {question}\n\n Answer: {predicted answer}
	Retrieval-augmented	Given the following information: \n\n {context}\n\n Can you judge if the following answer about the question is correct based on the given information or your internal knowledge, if yes, you should answer True or False, if no, you should answer "Unknown".\n\n Question: {question} \n\n Answer: {predicted answer}

Table 6: Prompt design for different perspectives of judgemental prompting with various settings.

B Supplement Analyses

B.1 Analysis on Retrieval Sources

Using supporting documents with high-efficiency contents can effectively improve the QA performance of LLMs. We conduct the analysis on three retrieval sources and utilize various metrics related to retrieval and QA in Figure 4 to thoroughly analyze the impact of retrieval augmentation from multiple perspectives. Here, the efficiency of documents can be evaluated through three factors: the presence of positive content (recall rate), proportion of positive documents in supporting documents (positive percentage), and token-level text density of positive

content (answers per 1K token). Firstly, the recall rate of documents generated by ChatGPT was significantly lower than that of the dense retriever. However, the QA performance using supporting documents generated by ChatGPT as supporting documents do not lag far behind dense retrieval. Additionally, LLaMA2's retrieval-augmented QA performance with ChatGPT as the retriever exceeds that where the dense retriever is implemented. We observe that among the documents from the three retrieval sources, dense retrieval results have the highest average proportion of positive documents (containing the correct answers), but the number of answers per token was not as high as in documents generated by ChatGPT. This phenomenon stems

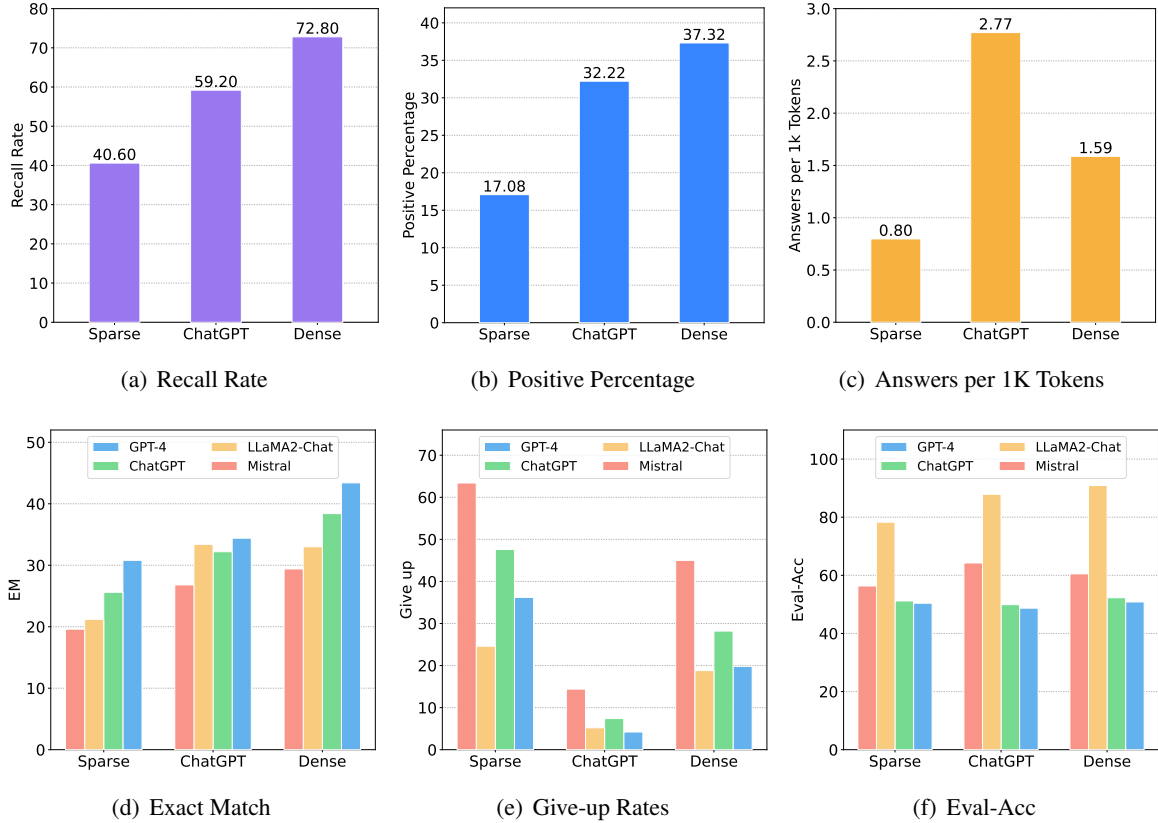| (a) Recall Rate | (b) Positive Percentage | (c) Answers per 1K Tokens |
| (d) Exact Match | (e) Give-up Rates | (f) Eval-Acc |

Figure 4: Metrics related to retrieval, QA, and judgement using different retrievers.

from the fact that documents generated by ChatGPT are more concise and the documents are closely related to the given questions. Although Wikipedia documents from the dense retriever have a higher recall rate and proportion of positive documents, they are relatively longer and contain more irrelevant information. As discussed earlier, this could be one of the reasons limiting the upper bound of retrieval augmentation. Furthermore, it could potentially contribute to the heightened confidence levels observed in LLMs when utilizing ChatGPT-generated documents as supporting documents, as evidenced by their low *Give-up* rates. In addition, the impact of different retrieval sources on *Eval-Acc* is not substantial.

B.2 Analysis on Query Types

Retrieval augmentation can change the query category preference of LLMs in QA and factual knowledge boundary perception. In order to investigate the propensity of LLMs to handle questions with varied characteristics, we separately calculate the answer accuracy of LLMs across different question categories. To achieve this, we utilize

supporting documents retrieved by the dense retriever. As shown in Figure 5, we can see that Chat-GPT and LLaMA2 achieve the highest *EM* when dealing with questions in the "*who*" and "*which*" category, indicating these types of questions may be the strong suit of the two LLMs. On the other hand, ChatGPT and LLaMA2 may not suffice for the question type of "*how*" in knowledge-intensive scenarios. When retrieval augmentation is incorporated, we observe that the preference of LLMs changes. The overall answer accuracies of LLMs are improved, and the accuracies in most categories increase proportionately. Specially, both ChatGPT and LLaMA2 perform best on the question type "*who*". We found that *EM* for certain question types does not increase with the overall EM increase (*e.g.*, "how" and "yesno" for ChatGPT, "which" for LLaMA2), indicating that retrieval augmentation cannot effectively enhance the QA performance for all question types. Figure 6 shows the Eval-Acc metric in each query type for ChatGPT and LLaMA2. We can see that most types of questions achieve better posteriori judgement accuracies, except "*how*" and "*yesno*" for LLaMA2 and "*yesno*" for ChatGPT. Similarly, the result indicates
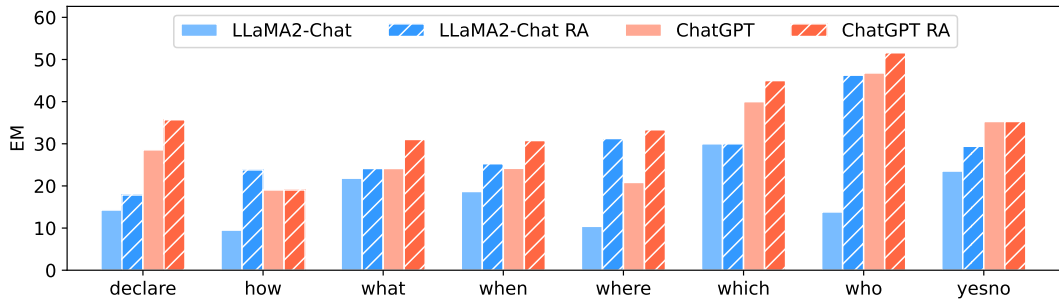
3712

Figure 5: The proportion of questions answered correctly by LLMs in different question categories under two QA prompting settings on NQ.
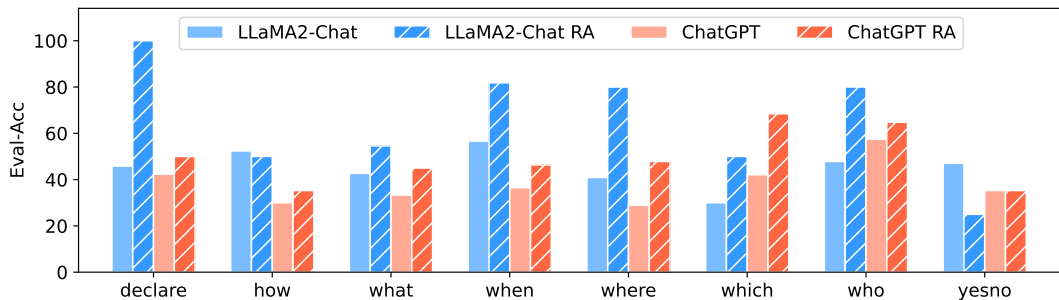

Figure 6: The proportion of answers judged correctly by LLMs in different question categories under two QA prompting settings on NQ.

that retrieval augmentation does not uniformly improve the knowledge boundary perception abilities of LLMs across all question types.

B.3 Analysis on Parameter Scale

Retrieval augmentation is more pronounced for improving LLMs with fewer parameters, including both QA performances and accuracies of knowledge boundary perception. Figure 7 illustrates the comparison of *EM* and *Eval-Acc* before and after retrieval augmentation for LLaMA2 models of varying parameter scales. We can observe that retrieval augmentation reduces the performance gaps in QA performance (shown by *EM*) of LLMs with different parameter scales, LLaMA2-7B model has the biggest improvement of *EM*. It is evident that retrieval augmentation mitigates the performance gaps in QA performance, as demonstrated by *EM* across LLMs with varying parameter scales. Notably, the LLaMA2-7B exhibits the most substantial improvement in *EM* than 13B and 70B models. Furthermore, the improvement in the accuracy of knowledge boundary perception of the LLaMA2-7B model with retrieval augmentation surpasses that of the 13B and 70B models, achieving the highest *Eval-Acc*. This phe-

nomenon further supports our speculation that in knowledge-intensive tasks, LLMs with smaller parameter scales possess less knowledge. Under a retrieval augmentation setting, these LLMs are likely to yield greater benefits compared to LLMs with larger parameter scales.

C Related Work

In this section, we review the related work from two perspectives, including large language models (LLMs) and open-domain question answering (QA).

C.1 Large Language Models

LLMs have significantly advanced the field of natural language processing (NLP), demonstrating remarkable abilities in generating human-like text and understanding complex language patterns (Zhao et al., 2023). These models are trained on vast corpora of text data and can perform a wide range of tasks without task-specific training (Thirunavukarasu et al., 2023; Lin et al., 2024; Ren et al., 2024a,b).

Typically, LLMs can fall into two categories: publicly available and closed source. Publicly available LLMs offer the research community the
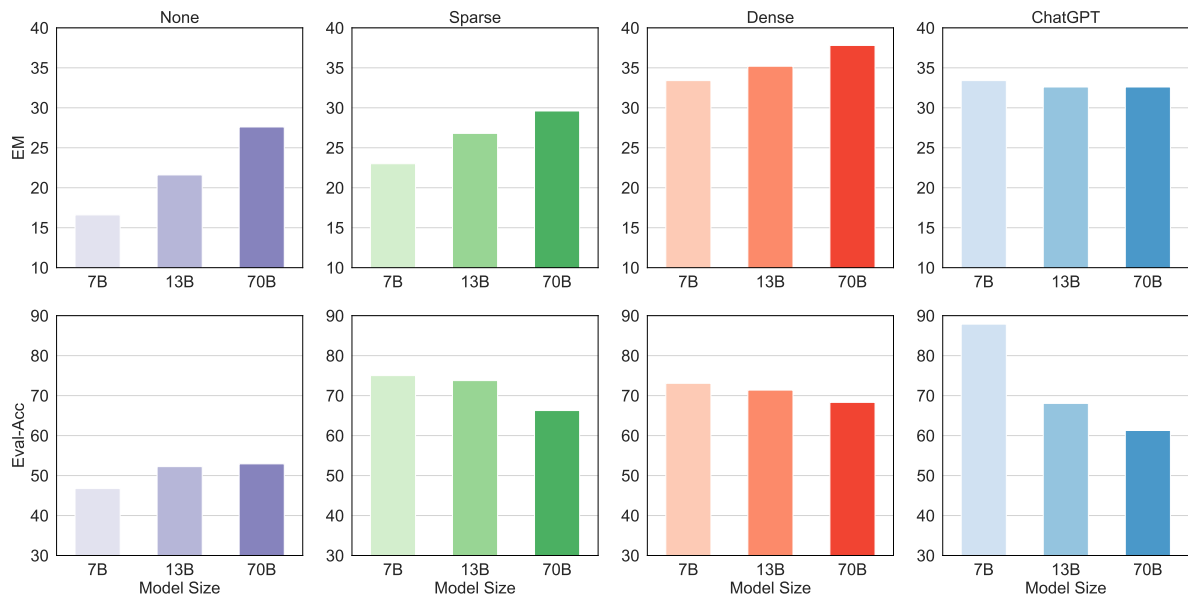
Figure 7: The performance of the LLaMA2 series, across its 7B, 13B, and 70B versions, showcases the influence of various retrieval document sources on *EM* and *Eval-Acc*.

flexibility to modify and adapt models to specific needs, fostering further innovation and collaboration. Among publicly available LLMs, the LLaMA series models stands out for its exceptional instruction-following performance (Touvron et al., 2023a,b), with ongoing efforts to fine-tune or continually pre-train its versions. Mistral, another open-source framework (Jiang et al., 2023a), provides a scalable solution for training LLMs on diverse datasets and supports various architectures to achieve top performance. Additionally, there are also LLMs like Falcon (TII, 2023), Vicuna (Chiang et al., 2023), and GLM (Zeng et al., 2022) actively contribute to advancing publicly available LLMs. Conversely, closed source LLMs, developed and maintained by private organizations, have non-public checkpoints and confidential internal workings. The typical closed source LLMs are GPT series proposed by OpenAI, including initial GPT to the latest GPT-4 (Radford et al., 2018; Brown et al., 2020b; OpenAI, 2023), have set new benchmarks in language understanding and generation, finding widespread use in commercial applications. There are also other closed source LLMs have been released, such as Sparrow (Glaese et al., 2022) and Claude. Some closed source LLMs are accessible via API calls, eliminating the need for local execution.

Despite the impressive advancements in LLMs, there remain unexplored boundaries regarding their knowledge and capabilities. Furthermore, the in-

corporation of retrieval augmentation in LLMs presents substantial potential for enhancing their knowledge, potentially reshaping the knowledge boundaries of LLMs. Therefore, it is necessary to comprehensively investigate the knowledge boundaries of LLMs and the effects of retrieval augmentation.

C.2 Open-domain Question Answering

Open-domain QA is a typical knowledge-intensive task in the field of NLP, aimed at providing precise and pertinent answers to questions posed in natural language, without the limitations of a specific domain or topic (Prager et al., 2007). This task is inherently challenging due to the vastness of human knowledge and the complexity of natural language understanding. Early approaches to open-domain QA rely heavily on structured knowledge bases, such as Freebase (Bollacker et al., 2008) or DBpedia (Auer et al., 2007), enabling precise yet limited retrieval of factual information. However, these approaches are constrained by the need for structured queries and the coverage of the underlying knowledge bases. The introduction of pre-trained language models, such as BERT (Devlin et al., 2019), BART (Lewis et al., 2020a), and their successors, marks a significant shift in the landscape of open-domain QA. These models leverage advanced text modeling techniques to understand natural language, enabling more flexible and context-aware answering performances.

3714

Recent advancements in open-domain QA have focused on incorporating external knowledge sources, such as Wikipedia or the web, to broaden the scope of answerable questions (Chen et al., 2017). This has led to the development of sophisticated information retrieval mechanisms (Ai et al., 2023), such as dense passage retrieval (DPR) (Karpukhin et al., 2020; Qu et al., 2021; Zhan et al., 2021; Zhuang and Zuccon, 2021; Zhan et al., 2022) and passage reranking (Nogueira and Cho, 2019; Nogueira et al., 2019; Gao et al., 2021), which efficiently extract and prioritize relevant information from large corpora (Zhao et al., 2022; Zhu et al., 2023; Guo et al., 2022). Additionally, techniques like few-shot learning and transfer learning have been explored to enhance the models' ability to generalize across diverse question types and domains, reducing the reliance on extensive labeled datasets (Brown et al., 2020a; Guu et al., 2020; Liu et al., 2022).

Despite significant progress, open-domain QA remains an active area of research, with ongoing efforts to improve answer accuracy (Izacard and Grave, 2021b; Jin et al., 2024), interpretability (Ribeiro et al., 2016), knowledge utilization (Wang et al., 2024; Zhou et al., 2024; Ni et al., 2024), and robustness to complex or ambiguous queries (Lewis et al., 2020b; Rajpurkar et al., 2018). Challenges such as handling out-of-domain questions, dealing with evolving knowledge, and ensuring fairness and transparency in answers continue to drive innovation in this field (Gebru et al., 2021; Mitchell et al., 2019). Furthermore, Several pioneering studies have been specifically designed to apply LLMs to open-domain QA tasks (Qin et al., 2023; Kamalloo et al., 2023; Yue et al., 2023; Wang et al., 2023a; Sun et al., 2023; Xu et al., 2024). Typically, they mainly focus on evaluating the QA performance of LLMs, discussing improved evaluation methods or leveraging LLMs to enhance existing open-domain QA models. Additionally, the existing effort also detects the uncertainty of LLMs with an automated method (Yin et al., 2023). As open-domain QA systems become increasingly integrated into real-world applications, they have potential to revolutionize access to information and knowledge.