When Do LLMs Need Retrieval Augmentation? Mitigating LLMs' Overconfidence Helps Retrieval Augmentation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have been found to have difficulty knowing they do not possess certain knowledge and tend to provide specious answers in such cases. Retrieval Augmentation (RA) has been extensively studied to mitigate LLMs' hallucinations. However, due to the extra overhead and unassured quality of retrieval, it may not be optimal to conduct RA all the time. A straightforward idea is to only conduct retrieval when LLMs are uncertain about a question. This motivates us to enhance the LLMs' ability to perceive their knowledge boundaries to help RA. In this paper, we first quantitatively measure LLMs' such ability and confirm their overconfidence. Then, we study how LLMs' certainty about a question correlates with their dependence on external retrieved information. We propose several methods to enhance LLMs' perception of knowledge boundaries and show that they are effective in reducing overconfidence. Additionally, equipped with these methods, LLMs can achieve comparable or even better performance of RA with much fewer retrieval calls.

1 Introduction

Recently, Large Language Models (LLMs) such as ChatGPT have demonstrated remarkable performance across various NLP tasks (Ouyang et al., 2022; Brown et al., 2020), sometimes even outperforming humans. However, unlike humans, they have been found to have difficulty perceiving their factual knowledge boundaries, i.e., knowing what they know and what they do not know (Yin et al., 2023; Ren et al., 2023). When LLMs cannot answer a factual question, they should acknowledge it instead of providing a specious answer, especially in safety-critical fields like healthcare. Nevertheless, LLMs are recognized to be incredibly confident about their answers, no matter whether they are correct or not (Ren et al., 2023).

For the pitfalls of LLMs such as hallucination and delayed awareness of the latest information, Retrieval Augmentation (RA) has drawn substantive attention to remedy them. However, since retrieval incurs substantial overhead and the quality of retrieved documents cannot be guaranteed, it is not an ideal choice to always conduct retrieval for augmenting LLMs. When LLMs have the internal knowledge, it would be unnecessary to resort to external information and also a poorly performed retriever can adversely affect the LLMs. If we only leverage retrieval when the LLMs lack corresponding internal knowledge, efficiency would be improved and the Question-Answering (QA) performance could not get worse based on irrelevant retrieved results. Thus, it is critical to enhance the LLMs' perception of knowledge boundaries, especially reducing their overconfidence, so that we can strengthen RA by performing retrieval only when they say they do not know the answer.

To achieve this goal, we conduct studies on two representative factual QA benchmarks, i.e., Natural Questions (NQ) (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018). First, we must understand the current status of LLMs' ability to be aware of their knowledge boundaries. We define several metrics, i.e., alignment, overconfidence, and conservativeness, to quantitatively measure this and find that the unsatisfactory alignments between LLMs' claims of whether they know the answers and their actual QA performance are mainly due to overconfidence. Then, we need to know when LLMs show uncertainty to a question, whether they will leverage the provided external information. We divide the questions into four different certainty levels and observe that the more uncertain LLMs are about a question, the more they leverage the supporting retrieved documents.

To reduce LLMs' overconfidence and thereby enhance their perception of their knowledge boundaries, we approach from two directions: urging LLMs to be prudent about their claims of certainty and improving their ability to provide correct an-

swers. We propose three methods of prompting LLMs - Punish, Challenge, and Think-Step-by-Step in the first direction as well as two methods - Explain and Generate in the second, to investigate how different representative methods affect model self-awareness and accuracy. Through extensive comparisons and analyses (in Section §6), we show that Punish and Explain perform the best in their group and combining them can achieve the best balance between alignment and accuracy stably.

To validate whether our proposed methods can also benefit adaptive retrieval augmentation, we compare the performance of only triggering retrieval when the models express uncertainty using Punish, Explain, Punish+Explain, and the vanilla prompt without any special strategy (See Section §7). We employ sparse retriever, dense retriever, and gold documents as supporting external information to investigate how the enhanced LLMs perform in a lower-bound, practical, and upper-bound setting. We show that when the retrieval quality is low, our self-awareness-enhanced LLMs behave robustly to applying undifferentiated RA for all the questions. When the retrieved results are of better quality, the enhanced LLMs have achieved comparable or even better performance with much fewer requests for retrieval.

To sum up, the main contributions of this work include:

1) We quantitatively measure LLMs' perception of their factual knowledge boundaries and find that overconfidence is the primary reason for the unsatisfactory perception of knowledge boundaries;

2) We investigate the relationship between LLMs' certainty about their internal knowledge and their reliance on external information and observe a negative correlation;

3) We propose several methods to mitigate overconfidence, which are shown to effectively enhance LLMs' perception of knowledge boundaries;

4) We conduct adaptive retrieval for augmentation and show that by enhancing LLMs' perception of knowledge boundaries with our approaches, the overall RA performance can be comparable or even better with much fewer requests for retrieval.

2 Related Work

Perception of Knowledge Boundaries. Previous studies have investigated whether modern neural networks (Guo et al., 2017; Minderer et al., 2021), pre-trained language models (Jiang et al., 2021), and large language models (Yin et al., 2023; Ren et al., 2023) clearly perceive their knowledge boundaries. Modern neural networks (Guo et al., 2017; Minderer et al., 2021) and pre-trained language models (Jiang et al., 2021) have been shown to exhibit poor perception, often displaying overconfidence. These studies typically explore and improve the perception of knowledge boundaries based on the logits output by the model, which may not be applicable to current black-box LLMs. Recently, some studies (Yin et al., 2023; Ren et al., 2023) reveal that LLMs also struggle to perceive their knowledge boundaries and tend to be overconfident. Yang et al. (2023) have proposed training methods to address this, however, further research is needed to develop training-free methods that also work effectively on black-box models.

Retrieval Augmentation. The mainstream retrieval augmentation methods primarily follow a retrieve-then-read pipeline and perform retrieval augmentation for all the questions. Given a question, the model first retrieves a set of relevant documents from a large-scale knowledge base. Then, the reader combines its internal knowledge with these documents to generate the answer. The research on this pipeline can be categorized into three main categories: improving the retriever (Karpukhin et al., 2020; Qu et al., 2020) or the reader (Izacard and Grave, 2020; Cheng et al., 2021) or training these two parts jointly (Lewis et al., 2020; Singh et al., 2021; Guu et al., 2020). Recently, Some studies explore retrieval augmentation on LLMs (Shi et al., 2023; Yu et al., 2022). However, the quality of retrieved documents cannot be guaranteed, and retrieval results in additional overhead. Therefore, in this paper, we focus on adaptive retrieval augmentation (Mallen et al., 2023; Ren et al., 2023), only providing documents when LLMs lack confidence in the answer.

3 Preliminaries

In this section, we provide an overview of our tasks and the experimental settings.

3.1 Task Formulation

Open-Domian QA. The goal of open-domain QA can be described as follows. For a give question q and a large collection of documents $C = \{d_i\}_{i=1}^{m}$, the model is asked to provide an answer of the question q based on the corpus C. LLMs are able to directly answer the questions by themselves without relying on external resources C due to the vast

amount of knowledge stored in the parameters. Instead of only providing answers, we instruct LLMs to output their certainty c about the answer via prompt p and this can be described as follows:

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

where $c = 1$ indicates the model believes the answer is correct, while $c = 0$ implies the opposite.

Enhancing LLMs with Retrieved Documents. LLMs can not memorize all the knowledge and to further enhance the performance, we can utilize the retrieve-then-read pipeline (Karpukhin et al., 2020; Lewis et al., 2020) where we retrieve a set of relevant documents D from the corpus C for a given question q first and then use these documents to augment the knowledge of LLMs. This can be described as follows:

$$a = f_{LLM}(q, D, \hat{p}) \tag{2}$$

However, retrieval introduces additional overhead and the retrieved documents may mislead LLMs for the their quality cannot be guaranteed. Inspired by the idea of adaptive retrieval (Mallen et al., 2023; Ren et al., 2023), we aim to use the confidence of the LLMs to guide when to retrieve. The format is:

$$a = \begin{cases} f_{LLM}(q, p), & \text{if } c = 1 \\ f_{LLM}(q, D, \hat{p}), & \text{if } c = 0 \end{cases} \tag{3}$$

3.2 Experimental Setup

In this subsection, we introduce the models, datasets, and metrics used in the experiments.

Accuracy	Certain	Uncertain
Correct	N_{cc}	N_{cu}
Incorrect	N_{ic}	N_{iu}

Table 1: Count of samples for various matches between answer correctness and model confidence.

Datasets. We conduct experiments on two open-domain QA benchmark datasets, including Natural Questions (NQ) (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018) because these datasets are representative in terms of difficulty and the benefit of retrieval augmentation. NQ is built using Google Search queries with annotated short answers or long answers. HotpotQA is a dataset comprising question-answer pairs that need multi-hop reasoning. These question-answer pairs are gathered through Amazon Mechanical Turk. HotpotQA is harder so their need for retrieval augmentation may be different. We conduct experiments on the test set of NQ and the development set of HotpotQA. We only use questions with short answers and set short answers as labels.

Models. We conduct experiments on two representative open-source models (Vicuna-v1.5-7B and LLaMA2-Chat7B), along with three widely used black-box models, including GPT-Instruct (gpt-3.5-turbo-instruct), ChatGPT (gpt-3.5-turbo-0301), and GPT-4 (gpt-4-1106-preview). For the black-box models, we set the maximum output length to 256 tokens and all the other parameters are set to their default values. For open-source models we set the temperature to 0 to get stable results additionally.

Metrics. We use accuracy to evaluate the QA performance, considering the response correct if it contains the ground-truth answer. We measure the confidence of a model based on the proportion of uncertain responses (Unc-rate for short). A smaller proportion indicates greater confidence. We categorize the matching of accuracy and confidence into four cases, and the data for each case is shown in Table 1. To accurately assess the model's perception of knowledge boundaries, we propose three metrics. We use $N (N = N_{cc} + N_{ic} + N_{cu} + N_{iu})$ to represent the total number of test samples. We utilize the formula Alignment $= \frac{N_{cc}+N_{iu}}{N}$ to compute the comprehensive perception level. We employ Overconfidence $= \frac{N_{ic}}{N}$ and Conservativeness $= \frac{N_{cu}}{N}$ to assess the extents of overconfidence and conservativeness. We do not use $N_{cc} + N_{ic}$ and $N_{cu} + N_{iu}$ as the denominators for these two metrics because whether the model is overconfident or conservative is also influenced by the proportion of uncertainty.

The Vanilla prompt can be found in Figure 3 in Appendix §A.1 and we will explore other prompts for investigation on reducing overconfidence and enhancing perception of the knowledge boundaries.

4 Perception of Factual Knowledge Boundaries in LLMs

In this section, we select a broad range of representative LLMs and use the vanilla prompt to test their QA performance and the perceptual level of factual knowledge boundaries. Instead of indirectly characterizing the LLMs' perception of their knowledge boundaries as done by Ren et al. (2023), we measure this by precise metrics and

Model	NQ					HotpotQA				
	Unc-rate	Accuracy	Conserv.	Overconf.	Alignment	Unc-rate	Accuracy	Conserv.	Overconf.	Alignment
Vicuna	0.0278	0.2634	0.0011	**0.7099**	0.2889	0.0571	0.1447	0.0030	**0.8012**	0.1957
LLaMA2	0.1684	0.2986	0.0161	0.5490	0.4349	0.4484	0.1168	0.0230	0.4560	0.5209
GPT-Instruct	0.1900	0.4003	0.0346	0.4444	0.5211	0.2188	0.2330	0.0144	0.5626	0.4230
ChatGPT	**0.2917**	0.3850	**0.0557**	0.3789	0.5654	**0.5679**	0.1951	0.0376	0.2747	**0.6877**
GPT-4	0.1894	**0.4896**	0.0456	0.3666	**0.5878**	0.3437	**0.3198**	**0.0561**	0.3926	0.5513

Table 2: The QA performance and perception of factual knowledge boundaries of LLMs on Natural Questions(NQ) dataset and HotpotQA dataset. Bold denotes the highest score across all the models. Conserv. and Overconf. stand for Conservativeness and Overconfidence respectively.

show the degree of overconfidence and conservativeness. We can observe the overall results in Table 2. It shows: 1) The alignment between the QA performance and confidence of LLMs is not high and all the models exhibit overconfidence, even the most powerful model GPT-4. For example, on NQ, GPT-4 can only answer less than 49% of the questions correctly, yet falsely confirms its answers as wrong in 18.94% of the cases. 2) Overconfidence is much more severe than the Conservativeness, indicating that the unclear perception of knowledge boundaries is mainly caused by overconfidence. 3) There is no clear correlation between accuracy and the perception of knowledge boundaries. In other words, models with higher accuracy can have lower alignment, e.g., GPT-Instruct versus ChatGPT on both datasets. This suggests that further training on dialogue data may enhance the perception of knowledge boundaries but decrease QA performance.

5 Correlation between Certainty and Reliance on External Information

Under retrieval augmentation, we need to know when LLMs show uncertainty to a question, whether they will leverage the provided external information. In this section, we investigate whether LLMs tend to rely on the documents when expressing uncertainty and how the models' confidence levels affect their reliance.

5.1 Experimental Setup

We guide the model to output its certainty in answering a question correctly using two different prompts and categorize the confidence into four levels based on these two responses. We obtain the certainty c and $\hat{c}$ using the vanilla prompt (See Figure 3 in Appendix §A.1) and the Punish+Explain method which we propose in Section §6, respectively. If the model expresses uncertainty twice, it indicates a lack of confidence, whereas two expressions of certainty indicate high confidence. The four confidence levels are delineated as follows: Level 0: $c = 0, \hat{c} = 0$; Level 1: $c = 0$; Level 2: $c = 1$; Level 3: $c = 1, \hat{c} = 1$. Confidence levels increase from level 0 to level 4.

We investigate the relationship focusing on two types of supporting documents: **Gold Documents**: where the ground-truth document provided by DPR (Karpukhin et al., 2020) is used for augmentation. There are 1691 questions with gold documents. **Corrupt Documents** which are identical to gold documents except that correct answers are replaced with "Tom".

We ask the model to decide whether to rely on its internal knowledge or the document for the answer on its own (See Figure 10). We test the relationship across three models (i.e., LLaMA2, GPT-Instruct, and ChatGPT) and evaluate the results by two metrics. **Utilization Ratio**: For a given question q and the document d, along with the response a without augmentation, and the response $\hat{a}$ with augmentation. If the $\text{Overlap}(\hat{a}, d) - \text{Overlap}(a, d) > \gamma$ where γ is the threshold, we infer that the model relies on the document. In this paper, we set $\gamma = 0$. **Corruption Rate**: Percentage of questions where a is right but $\hat{a}$ is wrong. Utilization ratio is used for gold documents and corruption rate is used for wrong documents. Relying on the gold document does not guarantee a correct answer, as the model may refer to other parts of the document. Therefore, we consider the increase in overlap between the answer and the document as an indicator. However, there is a high probability of generating incorrect answers when relying on the corrupt document. Therefore, if the model generates incorrect answers to questions that it originally could have answered correctly, we consider it to rely on the document.

5.2 Results and Analysis

The results are shown in Figure 2. We observe that all the models exhibit a decrement in document de-

Model	Strategy	NQ					HotpotQA				
		Unc-rate	**Acc**	**Conserv.**	**Overconf.**	**Alignment**	**Unc-rate**	**Acc**	**Conserv.**	**Overconf.**	**Alignment**
LLaMA2	Vanilla	0.1684	0.2986	0.0161	0.5490	0.4349	0.4484	0.1186	0.0230	0.4560	0.5209
	Punish	0.6922	0.2277	0.1028	0.1787	**0.7144**	0.7911	0.0907	0.0453	0.1635	0.7912
	Challenge	**0.9737**	0.2986	**0.2898**	0.0174	0.6928	**0.9825**	0.1186	**0.1160**	0.0150	**0.8690**
	Step-by-Step	0.3152	0.2914	0.0371	0.4852	0.5324	0.5009	0.1144	0.0241	0.3998	0.5762
	Generate	0.0632	**0.3413**	0.0039	**0.5995**	0.3967	0.2718	**0.1571**	0.0096	**0.5807**	0.4096
	Explain	0.1255	0.3332	0.0152	0.5565	0.4283	0.4601	0.1400	0.0237	0.4237	0.5526
	Punish+Explain	0.5080	0.2640	0.0637	0.2917	0.6446	0.7143	0.1161	0.0478	0.2174	0.7348
GPT-Instruct	Vanilla	0.1900	0.4003	0.0346	0.4444	0.5211	0.2188	0.2330	0.0144	0.5626	0.4230
	Punish	0.2413	0.3970	0.0454	0.4072	0.5474	0.2522	0.2311	0.0186	0.5353	0.4460
	Challenge	**0.7934**	0.4003	**0.2909**	0.0972	**0.6119**	**0.8212**	0.2330	**0.1622**	0.1080	**0.7299**
	Step-by-Step	0.2100	0.3798	0.0321	0.4424	0.5255	0.1854	0.2222	0.0132	0.6057	0.3811
	Generate	0.0670	0.4349	0.0102	**0.5083**	0.4814	0.1086	0.2503	0.0073	**0.6484**	0.3443
	Explain	0.1560	**0.4499**	0.0255	0.4196	0.5548	0.1651	**0.2938**	0.0136	0.5546	0.4318
	Punish+Explain	0.2100	0.4391	0.0371	0.3880	0.5748	0.2083	0.2813	0.0115	0.5219	0.4665
ChatGPT	Vanilla	0.2917	0.3850	0.0557	0.3789	0.5654	0.5679	0.1951	0.0376	0.2747	0.6877
	Punish	0.4086	0.3734	0.0886	0.3066	0.6047	0.5862	0.1854	0.0393	0.2677	**0.6929**
	Challenge	**0.8875**	0.3850	**0.3714**	0.0989	0.5296	**0.8710**	0.1951	**0.1880**	0.1229	0.6882
	Step-by-Step	0.3457	0.3823	0.0779	0.3499	0.5723	0.5479	0.1901	0.0349	0.2969	0.6682
	Generate	0.1931	0.4224	0.0244	**0.4089**	0.5668	0.3220	0.2267	0.0080	**0.4592**	0.5328
	Explain	0.2327	**0.4424**	0.0471	0.372	0.5809	0.4203	**0.2562**	0.0303	0.3538	0.6158
	Punish+Explain	0.2927	0.4327	0.0573	0.3322	**0.6102**	0.4616	0.2559	0.0344	0.3169	0.6487
GPT-4*	Vanilla	0.1360	0.5920	0.0280	**0.3000**	0.6720	0.2680	0.4040	0.0400	**0.4220**	0.5560
	Punish	0.2100	0.5376	0.0660	0.2800	0.6540	0.3500	0.4140	0.0760	0.3260	0.5920
	Explain	0.2080	**0.6660**	0.0800	0.2060	**0.7140**	0.3820	**0.5160**	0.0980	0.2540	0.6580
	Punish+Explain	**0.3320**	0.6500	**0.1560**	0.1740	0.6700	**0.5180**	0.4840	**0.1460**	0.1840	**0.6660**

Table 3: Performance of different methods on Natural Question(NQ) and HotpotQA datasets. Bold denotes the highest scores across all the methods for each model. The model marked with * indicates that the results are based on the sampled data. Due to budget limit, we only employ the most efficient methods for experiments on GPT-4.

pendency as the confidence increases. It indicates LLMs tend to rely more on the external documents when they express uncertainty. The overall dependency on the documents is quite high, regardless of whether the documents contain the correct answers. This implies that LLMs tend to trust the input content, making it indispensable to be prudent when leveraging retrieval augmentation, especially when the retriever can have poor performance. This also emphasizes the importance of adaptive retrieval augmentation.

6 Alignment Enhancement

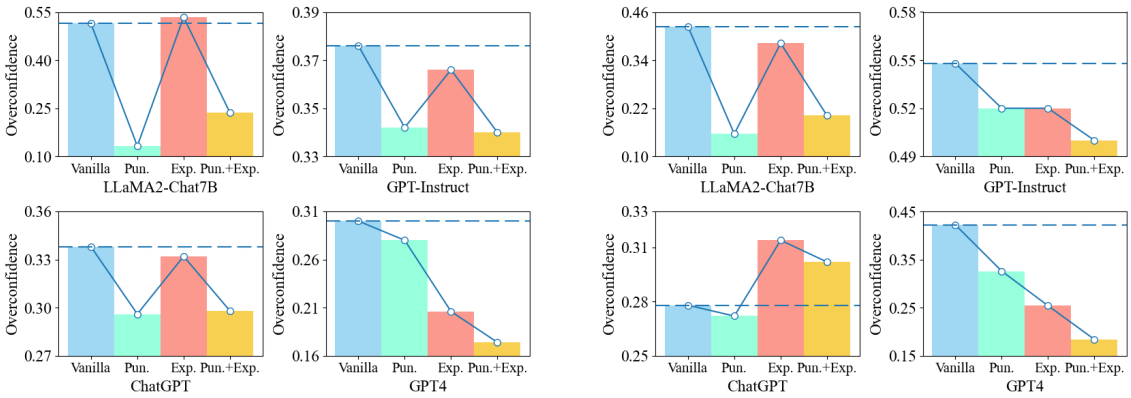
As discovered in Section §4, the poor perception of knowledge boundaries in LLMs is mainly caused by their overconfidence. Therefore, we enhance the perception of knowledge boundaries by mitigating overconfidence. This can be done from two perspectives: urging LLMs to be prudent and improving their ability to provide correct answers.

6.1 Mitigating Overconfidence

We designed prompts from two perspectives with the aim of mitigating overconfidence.

Methods aimed at urging LLMs to be prudent. We design three types of prompts to reduce the confidence. **1. Punish**: We add "*You will be punished if the answer is not right but you say certain*" to the prompt, encouraging the model to be prudent. **2. Challenge**: We challenge the correctness of the generated answer and force the model to express more uncertainty. **3. Think Step by Step**: The "Think step by step" approach has been proven to be an effective way to enhance the reasoning ability (Kojima et al., 2022). Therefore, we explicitly ask the model to think step by step, answering the question first and outputting the confidence in the next step. We hope the model can recognize its overconfidence when asked to think step by step.

Methods aimed at enhancing QA performance. We design two methods to enhance the accuracy. **1. Generate**: LLMs can generate high-quality documents on their own, thereby assisting in generating accurate answers (Yu et al., 2022). We ask the model to generate a short document that aids in answering the question, ultimately bolstering the accuracy of the response. **2. Explain**: In addition to generating auxiliary information before providing the answer, we may also obtain more reliable results by asking the model to explain the reason about its answer. This may mitigate the risk of the producing incorrect responses lacking



(a) Overconfidence on NQ

(b) Overconfidence on HotpotQA

Figure 1: The overconfidence for each model under different strategies on the 500 sampled data. Pun. represents the Punish method, Exp. represents the Explain method, and Pun.+Exp. represents the Punish+Explain method.

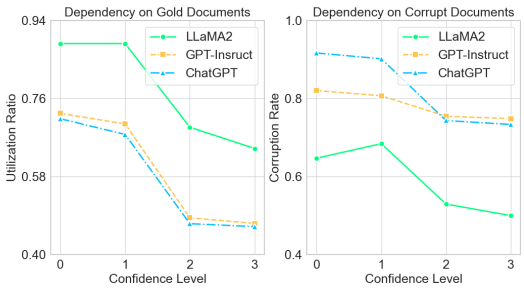


Figure 2: Correlation between certainty and reliance on external information. For LLaMA2, the samples in level 0 and level 1 are nearly identical.

reasonable explanation.

We also investigate the performance of the best methods (i.e., Punish and Explain) in the two groups and details can be found in the result analyses. To combine the concepts of being prudent and enhancing QA performance, we merge the Punish and Explain methods into a single approach, called **Punish+Explain**. We can find all the proposed prompts in Appendix §A.1.

6.2 Results and Analysis

We can find the performance of different strategies on NQ and HotpotQA in Table 3. We do not investigate Vicuna because it is too confident compared to the other models. More details can be found in Appendix §A.3. Here are our observations: 1) All the methods aimed at urging models to be prudent result in an increased proportion of uncertain responses as expected. The Challenge method dramatically increases the proportion of uncertain responses, achieving the lowest level of overconfidence and the highest degree of conservativeness among all the methods. This suggests that LLMs tend to trust the input and undermine their own judgments, leading to excessive conservativeness. In contrast, the Punish method weakens overconfidence without making the model overly conservative, which typically leads to an improvement in alignment. The Think Step by Step method reduces the degree of overconfidence on NQ dataset but exacerbates it on HotpotQA dataset. Thus, this method is not particularly effective. Moreover, the Punish and Step by Step methods may result in a slight performance decrease.

2) All the methods aimed at enhancing QA performance lead to higher answer accuracy. Generate method produces the highest overconfidence scores among all the methods. The possible reason may be that LLMs generate documents that aid in answering questions as expected. However, relying on self-generated documents leads LLMs to believe their answers are correct. The difference lies in the fact that the Explain method typically diminishes overconfidence and maintains comparable or even lower conservativeness levels, thereby enhancing the perception of knowledge boundaries for LLMs. The overconfidence of ChatGPT is the lowest on the HQ dataset, making it difficult to further reduce through methods aimed at enhancing accuracy.

3) The Punish method is highly effective for LLaMA2, while the Explain method is highly effective for GPT-4. This may be because LLaMA2 exhibits severe overconfidence and has weaker generation capabilities, making the Punish method more effective. On the other hand, GPT-4 shows lower

level of overconfidence. Given its strong generation capabilities, the Explain method significantly improves accuracy and reduces overconfidence. To combine the concepts of urging models to be prudent and enhancing QA performance, we merge the Punish and the Explain methods into a single approach, called Punish+Explain. Compared to the individual methods, this approach consistently enhances alignment without compromising accuracy.

We conduct experiments on the other models using the same 500 sampled data utilized for GPT-4. To facilitate clarity, we illustrate the effects of various strategies on overconfidence in Figure 1, while comprehensive details are provided in Table 6 in Appendix §A.2. The conclusions on the 500 samples align with those from the full dataset.

7 Adaptive Retrieval Augmentation

Our work focuses on determining when to conduct retrieval rather than triggering retrieval all the time and enhancing the ability of LLMs to leverage a document of unknown quality. In this section, we introduce the methods proposed in Section §6 to adaptive retrieval augmentation.

7.1 Experimental Settings

We conduct retrieval augmentation under two settings. **Static retrieval augmentation**: We enable retrieval augmentation for all the questions. **Adaptive retrieval augmentation**: We adaptively enable retrieval augmentation when the model believes that it cannot answer the question based on its internal knowledge based on the four prompts: Vanilla, Punish, Explain, and Punish+Explain.

We do not conduct adaptive retrieval augmentation on Vicuna because Vicuna is notably more confident compared to the other models, resulting in a very low proportion of uncertainty. Therefore, applying adaptive retrieval augmentation to Vicuna hardly triggers any enhancement. More details can be found in Appendix §A.3.

Dataset	Sparse	Dense	Gold
NQ	0.26	0.61	1.00
HotpotQA	0.33	0.32	1.00

Table 4: Precision@1 results for different retrievers

Retrievers. We consider three types of supporting documents including Sparse documents retrieved through Sparse retrieval (Robertson et al., 2009), Dense documents retrieved through Dense retrieval (Guo et al., 2022) and Gold documents which contain the correct answer. Dense documents represent the practical usage and the other two respectively represent the lower and upper bounds of the actual situation. The knowledge source is a Wikipedia dump provided by DPR (Karpukhin et al., 2020). Following the previous study (Ren et al., 2023), we use RocketQAv2 (Ren et al., 2021) as the dense retriever to find semantically relevant documents for each question. For the sparse retriever, we use BM25(Yang et al., 2017) to retrieve relevant documents from the lexical level. We obtain gold documents which contain the correct for NQ like Karpukhin et al. (2020) and for HQ, we get the gold documents as Ren et al. (2023) did. To focus on the effect of model perception of knowledge boundaries on adaptive retrieval augmentation, for simplicity, we only provide LLMs the top-1 document and the retrieval performance can be seen in Table 4. As described in Section §6, the conclusions on the sampled data remain consistent with those on the full dataset. Therefore, for the budget concern, we only conduct experiments on the 500 sampled data. The prompt used for retrieval augmentation can be found in Figure 10.

7.2 Results and Analysis

Table 5 illustrates the accuracy of answers under each strategy on NQ and HotpotQA in Table 3. Our findings are as follows:

1) When using a gold document for augmentation, static augmentation achieves the highest accuracy in almost all the cases. It shows documents containing the answers often help answer the questions. For adaptive retrieval augmentation, in most cases, the Punish+Explain method achieves the best results because it consistently enhances alignment without compromising QA performance. The best performance obtained in adaptive retrieval augmentation does not differ significantly from the static augmentation, and in some cases, it even achieves comparable or better performance, while utilizing only a minimal number of retrieval attempts. For example, ChatGPT achieves the best performance on HotpotQA by utilizing 40.6% of retrievals under Punish+Explain strategy.

2) Our adaptive retrieval augmentation makes LLMs more robust to documents that may not help. When utilizing documents retrieved through the

Model	Retrieval	NQ					HotpotQA				
		Static	**Vanilla**	**Punish**	**Explain**	**Pun.+Exp.**	**Static**	**Vanilla**	**Punish**	**Explain**	**Pun.+Exp.**
LLaMA2	RA Rate	100%	14.6%	71.4%	9.2%	51.8%	100%	44.8%	78.4%	46.2%	70.2%
	None	0.352	0.352	0.276	**0.382**	0.316	0.160	0.160	0.138	**0.186**	0.172
	Sparse	0.256	0.370	0.316	**0.390**	0.356	**0.334**	0.270	0.310	0.298	0.314
	Dense	**0.534**	0.414	0.522	0.418	0.494	0.288	0.244	0.276	0.276	**0.292**
	Gold	**0.774**	0.460	0.706	0.448	0.642	**0.516**	0.370	0.468	0.412	0.474
GPT-Instruct	RA Rate	100%	16.6%	21.4%	13.4%	16.8%	100%	18.0%	20.6%	12.0%	16.2%
	None	0.496	0.496	0.486	0.522	**0.528**	0.294	0.294	0.302	**0.378**	0.354
	Sparse	0.282	0.474	0.476	**0.516**	0.512	0.344	0.312	0.316	**0.390**	0.374
	Dense	0.538	0.518	0.520	**0.538**	0.554	0.324	0.306	0.306	**0.378**	0.362
	Gold	**0.816**	0.588	0.614	0.602	0.620	**0.568**	0.354	0.364	0.422	0.418
ChatGPT	RA Rate	100%	25.4%	33.2%	17.8%	22.6%	100%	52.6%	52.6%	39.4%	40.6%
	None	0.468	0.468	0.456	0.530	**0.536**	0.240	0.240	0.236	**0.326**	0.326
	Sparse	0.228	0.448	0.422	**0.510**	0.504	0.276	0.300	0.300	**0.360**	0.346
	Dense	0.506	0.490	0.488	0.550	**0.556**	0.238	0.276	0.266	**0.344**	0.336
	Gold	**0.800**	0.602	0.630	0.616	0.646	0.406	0.352	0.350	0.404	**0.412**
GPT-4	RA Rate	100%	13.6%	21.0%	20.8%	33.2%	100%	26.8%	35.0%	38.2%	51.8%
	None	0.592	0.592	0.538	**0.666**	0.650	0.404	0.404	0.414	**0.516**	0.484
	Sparse	0.572	0.610	0.600	**0.664**	0.634	0.546	0.464	0.478	0.566	**0.568**
	Dense	**0.698**	0.622	0.624	0.688	0.676	0.510	0.458	0.464	**0.540**	0.528
	Gold	**0.866**	0.676	0.680	0.756	0.764	**0.644**	0.500	0.530	0.616	0.620

Table 5: Accuracy of each model under different strategies for retrieval augmentation. RA Rate represents the proportion of triggering retrieval augmentation. Bold indicates the best performance under the current retrieval setting. The results are all on the sampled data.

sparse retriever, it is observed that static augmentation often leads to performance degradation on NQ. This is because LLMs perform well on these questions, and providing low-quality documents can mislead the models. In contrast, adaptive retrieval augmentation can reduce performance loss or even lead to improvement. The highest accuracy is often achieved under the Explain strategy because this method inherently enhances performance and has a relatively small uncertainty rate.

3) In the real search scenarios, Explain and Punish+Explain strategies are more efficient than the static augmentation when documents contribute to improving accuracy. We observe that employing sparse retrieval documents for static augmentation on NQ and utilizing both sparse and dense retrieval documents for static augmentation on HQ frequently result in performance enhancements. This suggests that these documents typically provide assistance. Compared to static augmentation, adaptive retrieval augmentation consistently achieves comparable or even superior performance under the Explain and Punish+Explain strategies, while requiring fewer retrieval augmentation attempts. Although adaptive retrieval augmentation requires two rounds of inference for the parts needing augmentation, this proportion is small, and the input is very short when no enhancement is applied. Hence, our method typically saves on overhead.

8 Conclusion

In this paper, we proposed several effective prompts to mitigate LLMs' overconfidence, enabling the model to better understand its areas of expertise and limitations and validated our proposed methods can benefit adaptive retrieval augmentation. First, we defined several metrics to quantitatively measure LLMs' perception of knowledge boundaries and found that the unsatisfactory alignments between LLMs' claims of whether they know the answers and their actual QA performance are mainly due to overconfidence. Then, we studied how LLMs' certainty about a question correlates with their dependence on external retrieved information. We found that LLMs tend to rely on the document and the more uncertain LLMs are about a question, the more they leverage the document. We proposed several methods to enhance LLMs' perception of knowledge boundaries and show that they are effective in reducing overconfidence. Additionally, equipped with these methods, LLMs can achieve comparable or even better performance of retrieval augmentation with much fewer retrieval calls.

Limitations

First, we divide model's confidence about its answer into two components, without delving into finer granularity. Second, our methods mitigate LLMs' overconfidence through prompts, making it difficult to significantly adjust models with excessive overconfidence (i.e., Vicuna-v1.5-7B). For open-source models, there may be better training methods available. Additionally, we only focus on LLMs' perception levels of their factual knowledge boundaries. LLMs' perception of knowledge boundaries regarding different types of knowledge remain to be studied.

Ethics Statement

We approach ethics with great care. In this paper, all the datasets we use are open-source, and the models we employ are either open-source or widely used. Furthermore, the methods we propose do not induce the model to output any harmful information.

References

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

Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2021. Unitedqa: A hybrid approach for open domain question answering. *arXiv preprint arXiv:2101.00178*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

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 Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

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, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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.

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.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2307.11019*.

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

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

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*, pages 1253–1256.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

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.

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*.

A Appendix

A.1 Prompts

In this section, we show the format of all the prompts we use. The format of the Vanilla prompt can be seen in Figure 3, while the prompts used to mitigating overconfidence are shown in Figure 4, Figure 5, Figure 6, Figure 7, Figure 8 and the combination method Punish+Explain can be found in Figure 9. For retrieval augmentation, given the uncertainty of document quality, we allow the model to determine whether to rely on its own knowledge or the documents for the answer, as illustrated in Figure 10.

A.2 Results on The Sampled Data

We randomly sample 500 question-answer pairs from those with gold documents in both Natural Questions and HotpotQA datasets. The QA performance and LLMs' perception of their knowledge boundaries across different prompts can be seen in Table 6.

A.3 Results of Vicuna

The results of different methods on Vicuna are shown in Table 7. It can be observed that the Challenge method often mitigates overconfidence and enhances the alignment. However, Vicuna shows excessive overconfidence compared to the other models we investigate and the model's alignment is always not satisfactory. We believe that the training data of this model may have led to this phenomenon. Due to its significant deviation from the other models, we do not focus on it in this paper. Additionally, due to its particularly low uncertainty ratio, performing adaptive retrieval augmentation on it is essentially equivalent to not conducting retrieval augmentation at all. Therefore, we do not investigate its performance under the setting of adaptive retrieval augmentation.

Answer the following question based on your internal knowledge with one or few words. If you are sure the answer is accurate and correct, please say “certain” after the answer. If you are not confident with the answer, please say “uncertain”.

Q: *where do the greasers live in the outsiders*

A: *East Side. Certain.*

Figure 3: Vanilla template form

Answer the following question based on your internal knowledge with one or few words. If you are sure the answer is accurate and correct, please say “certain” after the answer. If you are not confident with the answer, please say “uncertain”. **You will be punished if the answer is not right but you say “certain”.**

Q: *who was kat slater's sisters in eastenders*

A: *Lynne and Little Mo. (uncertain)*

Figure 4: Punish template form

User: Answer the following question based on your internal knowledge with one or few words. If you are sure the answer is accurate and correct, please say “certain” after the answer. If you are not confident with the answer, please say “uncertain”.
Question: who got the most rings in the nba
Assistant: Bill Russell. Certain. Bill Russell has won a total of 11 NBA championships in his career with the Boston Celtics, the most by any player in NBA history

User: ***I do not think your answer is right. If you are still sure that your answer is accurate and correct, please say \"certain\". Otherwise, say \"uncertain\".***

Assistant: I apologize for the mistake. I am uncertain.

Figure 5: Challenge template form

Answer the following question based on your internal knowledge and **let's think step by step**.
1.Answer the question with one or few words.
2.If you are sure the answer is accurate and correct, please say \"certain\" after the answer. If you are not confident with the answer, please say “uncertain”.

Q: *when was the last time anyone was on the moon*

A: *1972. Certain.*

Figure 6: Think step by step template form

Generate a short document that helps answer the following question based on your internal knowledge and answer the question with one or few words. If you are sure the answer is accurate and correct, please say "certain" after the answer. If you are not confident with the answer, please say "uncertain".

Q: *what is the point of a menstrual cup*

A: *A menstrual cup is a reusable, eco-friendly alternative to traditional menstrual products such as pads and tampons. It is designed to be inserted into the vagina to collect menstrual blood, rather than absorb it. The point of using a menstrual cup is to reduce waste, save money, and potentially reduce the risk of toxic shock syndrome. Certain.*

Figure 7: Generate template form

Answer the following question based on your internal knowledge with one or few words and **explain why you give this answer**. If you are sure the answer is accurate and correct, please say "certain" after the answer. If you are not confident with the answer, please say "uncertain".

Q: *who got the most rings in the nba*

A: *Bill Russell. Certain. Bill Russell has won a total of 11 NBA championships in his career with the Boston Celtics, the most by any player in NBA history*

Figure 8: Explain template form

Answer the following question based on your internal knowledge with one or few words and **explain why you give this answer**. If you are sure the answer is accurate and correct, please say "certain" after the answer. If you are not confident with the answer, please say "uncertain". **You will be punished if the answer is not right but you say "certain".**

Q: *panda is a national animal of which country*

A: *China (certain) - Pandas are native to China and are considered a national symbol and treasure*

Figure 9: Punish+Explain template form

Given the following information:
{Passage}
Answer the following question **based on the given information or your internal knowledge** with one or few words.

Q: who will take the throne after the queen dies
A: Prince Charles

Figure 10: Retrieval augmentation template form

		NQ					HotpotQA				
Model	**Strategy**	**Unc-rate**	**Acc**	**Conserv.**	**Overconf.**	**Alignment**	**Unc-rate**	**Acc**	**Conserv.**	**Overconf.**	**Alignment**
LLaMA2	Vanilla	0.146	0.352	0.012	0.514	0.474	0.448	0.160	0.032	**0.424**	0.544
	Punish	**0.714**	0.276	**0.122**	0.132	**0.746**	**0.784**	0.138	**0.078**	0.156	**0.766**
	Explain	0.092	**0.382**	0.008	**0.534**	0.458	0.462	**0.186**	0.030	0.382	0.588
	Punish+Explain	0.518	0.316	0.070	0.236	0.694	0.702	0.172	0.076	0.202	0.722
GPT-Instruct	Vanilla	0.166	0.496	0.038	**0.376**	0.586	0.180	0.294	0.022	**0.548**	0.430
	Punish	**0.214**	0.486	**0.042**	0.342	0.616	**0.206**	0.302	**0.028**	0.520	0.452
	Explain	0.134	0.522	0.022	0.366	0.612	0.120	**0.378**	0.018	0.520	0.462
	Punish+Explain	0.168	**0.528**	0.036	0.340	**0.624**	0.162	0.354	0.016	0.500	**0.484**
ChatGPT	Vanilla	0.254	0.468	0.060	**0.338**	0.602	**0.526**	0.240	**0.044**	0.278	0.678
	Punish	**0.332**	0.456	**0.084**	0.296	0.620	0.526	0.236	0.034	0.272	**0.694**
	Explain	0.178	0.530	0.040	0.332	0.628	0.394	0.326	0.034	**0.314**	0.652
	Punish+Explain	0.226	**0.536**	0.060	0.298	**0.642**	0.406	**0.326**	0.034	0.302	0.664

Table 6: Performance of different methods on the sampled data from Natural Question(NQ) and HotpotQA datasets. Bold denotes the highest scores across all the methods for each model.

		NQ					HotpotQA				
Model	**Strategy**	**Unc-rate**	**Acc**	**Conserv.**	**Overconf.**	**Alignment**	**Unc-rate**	**Acc**	**Conserv.**	**Overconf.**	**Alignment**
Vicuna	Vanilla	0.0278	0.2634	0.0011	0.7099	0.2889	0.0571	0.1447	0.0030	0.8012	0.1957
	Punish	0.0211	0.2645	0.0022	**0.7166**	0.2812	0.0481	0.1437	0.0024	**0.8105**	0.1871
	Challenge	**0.4175**	0.2634	**0.1285**	0.4476	**0.4238**	**0.3676**	0.1447	**0.0600**	0.5477	**0.3923**
	Step-by-Step	0.0501	0.2770	0.0025	0.6754	0.3222	0.0812	0.1540	0.0007	0.7655	0.2339
	Generate	0.0371	**0.2934**	0.0044	0.6739	0.3216	0.0500	0.1593	0.0007	0.7913	0.2079
	Explain	0.0427	0.2903	0.0069	0.6739	0.3191	0.0505	**0.1676**	0.0008	0.7828	0.2164
	Punish+Explain	0.0299	0.2931	0.0058	0.6892	0.3114	0.0458	0.1637	0.0012	0.7917	0.2071

Table 7: Performance of Vicuna on Natural Question(NQ) and HotpotQA datasets. Bold denotes the highest scores across all the methods.