Internal and External Knowledge Interactive Refinement Framework for Knowledge-Intensive Question Answering

Anonymous ACL submission

Abstract

Recent works have attempted to integrate external knowledge into LLMs to address the limitations and potential factual errors in LLM-generated content. However, how to retrieve the correct knowledge from the large amount of external knowledge imposes a challenge. To this end, we empirically observe that LLMs have already encoded rich knowledge in their pretrained parameters and utilizing these internal knowledge improves the retrieval of external knowledge when applying them to knowledge-intensive tasks. In this paper, we propose a new internal and external knowledge interactive refinement paradigm dubbed IEKR to utilize internal knowledge in LLM to help retrieve relevant knowledge from the external knowledge base, as well as exploit the external knowledge to refine the hallucination of generated internal knowledge. By simply adding a prompt like "Tell me something about" to the LLMs, we try to review related explicit knowledge and insert them with the query into the retriever for external retrieval. The external knowledge is utilized to complement the internal knowledge into input of LLM for answers. We conduct experiments on 3 benchmark datasets in knowledge-intensive question answering task with different LLMs and domains, achieving the new state-of-the-art. Further analysis shows the effectiveness of different modules in our approach.

1 Introduction

Large Language Models (LLMs) have shown remarkable abilities for human language processing and extraordinary scalability and adaptability in few- or zero-shot settings (Brown et al., 2020; Ouyang et al., 2022; Chowdhery et al., 2023). In spite of LLM's ability to generate plausible-sounding text, hallucination can occur when the model produces text that includes facts or claims that are fictional or misleading rather than providing reliable and truthful information (Yao et al., 2022; Bang et al., 2023). To mitigate these limitations, recent works propose retrieval augmented generation (RAG) (Lewis et al., 2020; Izacard et al., 2022; Khattab et al., 2022) to integrates external knowledge retrieved into the generative process. However, there is inevitably a gap between the input text and the needed knowledge in retrieval (Ma et al., 2023) and how to retrieve the right knowledge remains a challenge. To this end, we empirically discover the internal knowledge encoded in LLM parameters help the retriever to obtain the correct external knowledge in demand for knowledge-intensive tasks.

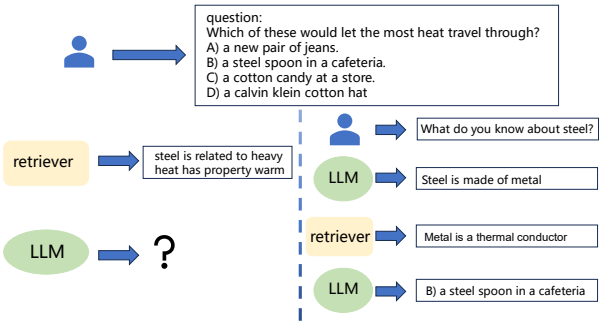

Figure 1: One example from OpenbookQA dataset

In Figure 1, the query is "Which of these would let the most heat travel through? A) a new pair of jeans. B) a steel spoon in a cafeteria. C) a cotton candy at a store. D) a calvin klein cotton hat", and the needed knowledge to answer the question is "Metal is a thermal conductor". There is a gap between the option in query "steel" and this external knowledge, so the retriever chooses many relevant but not needed knowledge like "steel is related to heavy, heat has property warm" which distracts the LLM to answer the question. However, the knowledge gap can be filled by prompt the LLM to reflect on its internal knowledge about the steel.

To improve the retrieval of needed external knowledge for question answering (QA), we pro-

pose our internal and external knowledge interactively refinement framework (IEKR), where the internal knowledge within LLM is utilized to retrieve needed knowledge in external knowledge base (KB), and the external knowledge retrieved is incorporated into complementing the internal knowledge. Specifically, first we prompt the LLM to generate the intrinsic knowledge about the concepts in the query. Then we input the internal knowledge along with the query to the language model (LM) retriever and get the top-k knowledge sentences from the external KB. The internal and external knowledge is inputted to the reader to answer the question. We conduct experiments on three benchmark knowledge-intensive QA datasets across different domains, OpenbookQA, CommonsenseQA and MedQA, becoming the new state-of-the-art. Further experimental results demonstrate the effectiveness of internal knowledge to retrieve the needed external knowledge, as well as the complement of internal knowledge with external knowledge.

To conclude, we summarize the contributions of this work as follows: **1.** We explicitly incorporate internal knowledge within LLM to retrieve needed external knowledge for knowledge-intensive QA, filling the gap between the input text and the needed knowledge in retrieval. **2.** We introduce self-criticism to refine the internal knowledge based on the retrieval and highlight the reflection process for further reasoning. **3.** We introduce self-conclude to conclude the intermediate result and provide clues for further retrieval. **4.** We conduct experiments on three benchmark datasets with different LLMs across different domains, and derive the SOTA performance. Further experiments show the effectiveness of different modules of our approach.

2 Related Work

Previous studies have shown that PLMs implicitly contain a large amount of knowledge. Petroni et al. (2019) have shown that such language models can be used in a KB completion task by converting KB relations into natural language templates. Based on this finding, researchers attempt to treat the PLM as a knowledge base. Some studies (Bosselut et al., 2019; West et al., 2021) employ PLMs to construct knowledge graphs automatically. Meanwhile, some others (Shwartz et al., 2020; Li et al., 2022) find that the knowledge possessed by the PLMs can be used to enhance the model's performance in downstream tasks. To date, several work (Wang et al., 2022; Zelikman et al., 2022) attempt to utilize PLMs to generate free-text rationales for reasoning. Our approach differs from previous works in that we aim to utilize the internal knowledge in LLMs to enhance the external knowledge retrieval.

Using interactive question-knowledge alignment, Zhang et al. (2023) presents a method for mitigating language model hallucination Their proposed approach focuses on aligning generated text with relevant factual knowledge, enabling users to interactively guide the model's responses to produce more accurate and reliable information. This technique aims to improve the quality and factuality of language model outputs by involving users in the alignment process. LLMAUGMENTER (Peng et al., 2023) improves LLMs using external knowledge and automated feedback. It highlights the need to address the limitations and potential factual errors in LLM-generated content. This method involves incorporating external knowledge sources and automated feedback mechanisms to enhance the accuracy and reliability of LLM outputs. By doing so, the paper aims to mitigate factual inaccuracies and improve the overall quality of LLM-generated text. Similarly, Li et al. (2023) introduces a framework called "Chain of Knowledge" for grounding LLMs with structured knowledge bases. Grounding refers to the process of connecting LLM-generated text with structured knowledge to improve factual accuracy and reliability. This approach aims to improve the alignment of LLM-generated content with structured knowledge, reducing the risk of generating inaccurate or hallucinated information. These methods neglect the internal knowledge within the LLM and there remains a gap between between the input text and the needed knowledge in retrieval.

3 Task Formulation

We focus on the multi-choice QA task (Robinson et al., 2022). The query includes a natural text question and several candidates and the model needs to choose an answer from the candidates:

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a} \in \mathcal{C}} P(\mathbf{a} \mid \mathbf{q}) \quad (1)$$

where $\mathcal{C}$ denotes the answer candidates and q denotes the question. There is an external KB to provide external knowledge for the model. The KB contains a series of factual triples and each triple (fact) is composed of two entities and one relation.

A KB can be denoted as $\mathcal{G} = \{(e, r, e')|e, e' \in E, r \in R\}$, where G denotes the KB, E denotes the entity set and R denotes the relation set.

4 Methodology

In this part, we introduce the architecture of our approach. Our model composes 4 steps: internal reflection, external retrieval, self-criticism and self-conclusion.

For an input query, first we prompt the LLM $\mathcal{M}$ to reflect on its internal knowledge about the query entities. Secondly we utilize the internal knowledge as well as query to retrieve the relevant external knowledge by retriever $\mathcal{R}$ to complement internal knowledge. Then the model critic the internal knowledge based on the complementary external knowledge. Then the model conduct self-conclude to derive intermediate reasoning result.

4.1 Internal Knowledge Reflection

In this part, we aim to dig into the LLM about the internal knowledge about the entities in query. We utilize off-shelf Named Entity Recognition model to extract the named entities in the query. For example, as to "Frilled sharks and angler fish live far beneath the surface of the ocean", we extract the "Frilled sharks", "angler fish" and "ocean" as the named entities.

For each entity, we construct the prompt like "tell me something about [entity]" to ask LLM $\mathcal{M}$ to generate the internal knowledge about it. We arrange the generation about each entity sequentially as the internal knowledge of LLM $\mathcal{M}$ of the current query:

$$\mathbf{ik_i} = \mathcal{M}(\mathbf{P} + \mathbf{e_i}) \quad (2)$$

$$\mathbf{IK} = \mathbf{Concate}[\mathbf{ik_1}||\mathbf{ik_2}||\cdots||\mathbf{ik_n}] \quad (3)$$

where $\mathbf{P}$ denotes the reflection prompt to LLM $\mathcal{M}$, n denotes the number of entities, $\mathbf{e_i}$ denotes the i-th entity in query.

4.2 External Knowledge Retrieval

In this part, we utilize a pretrained LM to retrieve the relevant knowledge from the external KB $\mathcal{G}$ base on the internal knowledge IK and query as the complement. The KB contains a large set of triples of the form (h, r, t), like (ice, HasProperty, cold), where h and t represent head and tail entities in the entity set V and r is a certain relation type from the pre-defined set. To reduce the search space, we following (Zhang et al., 2022) to perform entity linking to $\mathcal{G}$ to derive an initial set of nodes $\mathbf{V_{linked}}$ and prune the original KB $\mathcal{G}$ to keep the entities within 2 hops from $\mathbf{V_{linked}}$. We convert the triples into natural language templates, such as "(ice, HasProperty, cold)" is converted into "Ice has the property of cold".

The retriever contains a pretrained cross-encoder with transformer architecture, on top of which there is a classifer to predict the similarity score between the internal knowledge and the external knowledge sentence. We choose the top-k external sentences based on the similarity score as complements of internal knowledge:

$$\mathbf{s_i} = \mathbf{MLP}(\mathcal{R}([\mathbf{q} : \mathbf{IK} : \mathbf{ek_i}])) \quad (4)$$

$$\mathbf{EK} = \mathbf{Concate}[\mathbf{ek_{(1)}}||\mathbf{ek_{(2)}}||\cdots||\mathbf{ek_{(m)}}] \quad (5)$$

where $\mathbf{q}$ denotes the query, $\mathbf{ek_i}$ denotes the i-th external sentence, $\mathbf{ek_{(i)}}$ denotes the i-th external sentence after sorting by $\mathbf{s_i}$ and m denotes the number of external knowledge sentence.

4.3 Self-Criticism

In this part, we prompt the LLM to correct the internal knowledge based on the retrieved external knowledge. Considering the hallucination of LLM, there may be some incorrect or distracting knowledge in the generation of section 4.1. We input each piece of internal knowledge and the relevant external knowledge and ask the LLM to refine the correctness of internal knowledge:

$$\mathbf{ek_j} = \mathbf{argmax_t}\mathbf{MLP}(\mathcal{R}([\mathbf{ik_i}, \mathbf{ek_t}])) \quad (6)$$

$$\mathbf{c_i} = \mathcal{M}([P, \mathbf{ik_i} + \mathbf{ek_j}]) \quad (7)$$

$$\mathbf{Crit} = \mathbf{Concate}[\mathbf{c_1}||\cdots||\mathbf{c_n}] \quad (8)$$

where $\mathbf{ek_j}$ denotes the most relevant item of external knowledge about the i-th internal reflection $\mathbf{ik_i}$ and P denotes the respective prompt.

4.4 Self-Conclude

In this part, we prompt the model $\mathcal{M}$ to select the current helpful internal knowledge and external knowledge, and derive the intermediate result for the query. If the intermediate result does not derive the final answer, the model generates the required information such as "should figure out if medal is a thermal conductor and what calvin klein is" in Figure 2:

$$\mathbf{Con} = \mathcal{M}([P, \mathbf{IK}, \mathbf{EK}, \mathbf{Crit}]) \quad (9)$$

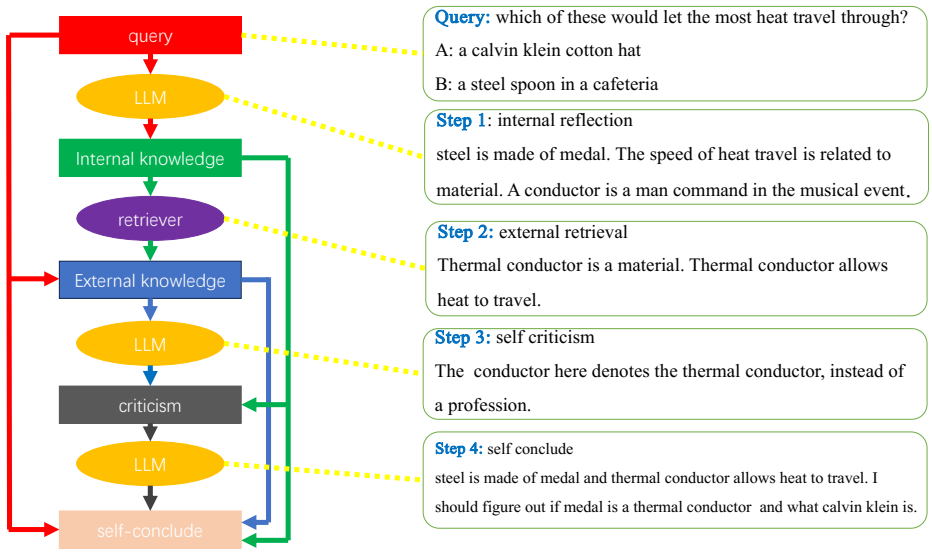

Figure 2: The pipeline of our approach. Our model composes 4 steps: internal reflection, external retrieval, self-criticism and self-conclusion. Red denotes the question inputted in respective step, and blue and green denote the external knowledge and internal knowledge respectively.

where **Con** denotes the generated self-conclusion and P denotes the repective prompt. We iteratively start the next phrase of internal reflection, external retrieval and self criticism based on the required information.

4.5 Answer Generation

In this part, we input the final internal knowledge, external knowledge, self-criticism as well as the self-conclusion into the LLM $\mathcal{M}$ to generate the answer. The cross-entropy loss is utilized to optimize the model:

$$L = -\log p(a|q, \mathbf{IK}, \mathbf{EK}, \mathbf{Crit}, \mathbf{Con}) \quad (10)$$

$$= -\log \prod_{i=1}^{l} p(a_i|q, \mathbf{IK}, \mathbf{EK}, \mathbf{Crit}, \mathbf{Con}, a_{<i}) \quad (11)$$

$$= -\sum_{i=1}^{l} \log p(a_i|q, \mathbf{IK}, \mathbf{EK}, \mathbf{Crit}, \mathbf{Con}, a_{<i}) \quad (12)$$

where a denotes the answer generated, l denotes the answer length, **Crit** denotes the self-criticism and **Con** denotes the self-conclusion .

5 Experiments

5.1 Datasets

We evaluate IEKR on three diverse multiple-choice question answering datasets across two domains: CommonsenseQA (Talmor et al., 2018) and OpenBookQA (Mihaylov et al., 2018) as commonsense reasoning benchmarks, and MedQA-USMLE (Jin et al., 2021) as a clinical QA task. CommonsenseQA is a 5-way multiple-choice question answering dataset of 12,102 questions that require background commonsense knowledge beyond surface language understanding. We perform our experiments using the in-house data split of (Lin et al., 2019) to compare to baseline methods. OpenbookQA is a 4-way multiple-choice question answering dataset that tests elementary scientific knowledge. It contains 5,957 questions along with an open book of scientific facts. We use the official data splits from (Mihaylov et al., 2018). MedQA-USMLE is a 4-way multiple-choice question answering dataset, which requires biomedical and clinical knowledge. The questions are originally from practice tests for the United States Medical License Exams (USMLE). The dataset contains 12,723 questions. We use the original data splits from (Jin et al., 2021).

5.2 Implementation Details

We utilize the widespread Flan-T5 (3B) (Chung et al., 2022) as the LLM $\mathcal{M}$ for CommonsenseQA and OpenbookQA. Because MedQA-USMLE requires more domain-specific knowledge, we choose the LLama2 (7B) (Touvron et al., 2023) as the LLM $\mathcal{M}$. We adopts the same verifier $\mathcal{V}$ for the three datasets, which is initialzed with LLama2 (7B). To reduce computation cost and keep prior

Methods	Dev	Para
LM+GNN		
MHGRN (Feng et al., 2020)	74.5	-
QA-GNN (Yasunaga et al., 2021)	76.5	-
GREASELM (Zhang et al., 2022)	78.5	-
RumiDeBERTa (Yao et al., 2023)	74.3	-
GrapeQA (Taunk et al., 2023)	74.9	-
Dragon (Yasunaga et al., 2022)	76.0	-
LLM		
GPT-3 (Xu et al., 2021b)	73.0	175B
UnifiedQA (Khashabi et al., 2020)	79.1	11B
Flan-T5 (Chung et al., 2022)	83.2	3B
RumiFlanT5 (Yao et al., 2023)	84.3	3B
GKP (Liu et al., 2021)	85.3	11B
LLM+RAG		
ReFeed	88.8	3B
FLARE	90.5	3B
Self-RAG	91.0	3B
Step-Back	90.7	3B
IEKR	**93.7**	3B

Table 1: Results on CommonsenseQA dataset compared with LM+GNN and LLM+RAG based methods. We adopt accuracy as the metric to evaluate the performance. GKP denotes Generated Knowledge Prompting. $para$ denotes the parameter number in the model.

knowledge in LLM, we use LoRA (Hu et al., 2021), which freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the LLM $\mathcal{M}$. So the number of trainable parameters of IEKR-7B is 4.5M, only 0.06% of total parameters of LLaMA2-7B.

We use ConceptNet (Speer et al., 2017), a general-domain knowledge graph, as our external knowledge source $\mathcal{G}$ for both CommonsenseQA and OpenbookQA. It has 799,273 nodes and 2,487,810 edges in total. For MedQA-USMLE, we use a self-constructed knowledge graph that integrates the Disease Database portion of the Unified Medical Language System (Bodenreider, 2004) and DrugBank (Wishart et al., 2018). The knowledge graph contains 9,958 nodes and 44,561 edges.

We adopt the pretrained dense retriever BGE-Reranker (Chen et al., 2024) to initialize the retriever $\mathcal{R}$. The number of cases N sampled from training dataset to finetune $\mathcal{R}$ is set to 500. The number of external knowledge sentences m retrieved by $\mathcal{R}$ wth the query and internal knowledge is set to 50. We utilize AdamW as the optimizer to train the LLM $\mathcal{M}$ and the learning rate is set to 3e-5. Training batch size is set to 2.

5.3 Baselines

We compare our methods with 2 groups of baselines: LM based methods with graph neural network (GNN) and LLM based methods with retrieval augmented generation (RAG). LM+GNN methods utilize GNN to incorporate the external knowledge from KB for knowledge-intensive QA. Because GNN involves much computation cost which does not apply to LLM, LLM+RAG based methods adopt retrieval augmented generation (RAG) to retrieve the external knowledge text and conduct task specific finetuning.

For LM+GNN methods, we compare our method with RoBERTa-Large RGCN (Schlichtkrull et al., 2018), GconAttn (Wang et al., 2019), KagNet (Lin et al., 2019), RN (Santoro et al., 2017), MHGRN (Feng et al., 2020), QA-GNN (Yasunaga et al., 2021), GREASELM (Zhang et al., 2022), RumiDeBERTa (Yao et al., 2023), GrapeQA (Taunk et al., 2023), and Dragon (Yasunaga et al., 2022) for OpenbookQA and CommonsenseQA; as well as SapBERT (Liu et al., 2020), QA-GNN (Yasunaga et al., 2021), GREASELM (Zhang et al., 2022), and GrapeQA (Taunk et al., 2023) for MedQA.

For LLM based methods, we compare our method with GPT-3 (Xu et al., 2021b), UnifiedQA (Khashabi et al., 2020), Flan-T5 (Chung et al., 2022), RumiFlanT5 (Yao et al., 2023), DeBERTa-xxlarge (Xu et al., 2021b), GKP (Liu et al., 2021), and GenMC (Huang et al., 2022) for OpenbookQA and CommonsenseQA; as well as GPT-Neo (Black et al., 2022), LLama2 (Touvron et al., 2023), and RumiLLama2 (Yao et al., 2023) for MedQA. For RAG based LLM, we compare with FLARE(Jiang et al., 2023) and ReFeed (Yu et al., 2023), and replace the original backbone from GPT-3.5 to the same as ours. We also compare with strong RAG methods self-RAG (Asai et al., 2023) and Step-back (Zheng et al., 2023).

5.4 Results

Our results in Tables 1 and **??** demonstrate a consistent improvement on the CommonsenseQA dataset compared with existing LM+GNN based methods and LLM-based methods, becoming the new state-of-the-art. Compared with competitive LM+GNN methods, GREASELM and Dragon, our model outperforms by 9.4 accuracy on dev set and 17.7 accuracy on in-house test set. Compared with best LLM based method, Generated Knowledge Prompting, our model outperforms by 2.6 accuracy on dev set.

In Tables 2, our improvements significantly outperforms the existing LM+GNN and LLM based methods on OpenbookQA dataset, becoming the new SOTA. IEKR improves the performance by 7.3 accuracy compared with the best LM+GNN method GREASELM, and 2.3 accuracy compared with the best LLM method GenMC on test set. It demonstrates the effectiveness of our internal and external knowledge interactively refinement framework.

Our reported results thus far demonstrate the viability of our method in the general commonsense reasoning domain. Further, we explore whether IEKR could be adapted to other domains by evaluating on the MedQA-USMLE dataset. Our results in Tables **??** and 3 demonstrate that IEKR outperforms SOTA LM+GNN method GrapeQA by 11.4 accuracy and LLM based method RumiLLama2 by 3.5 accuracy. It demonstrates that with different LLMs, our approach shows stable improvement over the knowledge intensive QA task across different domains. Our method effectively reflects on useful internal knowledge within the model, and utilize it to enhance the retrieval of external knowledge for QA task.

Methods	Test	Para
LM+GNN		
Dragon (Yasunaga et al., 2022)	72.0	-
RumiDeBERTa (Yao et al., 2023)	76.0	-
ALBERT+KG (Lan et al., 2019)	81.0	-
HGN (Yan et al., 2020)	81.4	-
AMR-SG (Xu et al., 2021a)	81.6	-
ALBERT+KPG (Wang et al., 2020)	81.8	-
DEKCOR (Xu et al., 2021c)	82.2	-
QA-GNN (Yasunaga et al., 2021)	82.8	-
GREASELM (Zhang et al., 2022)	84.8	-
LLM		
GPT-3 few shot (Xu et al., 2021b)	73.0	175B
T5 (Raffel et al., 2020)	83.2	3B
T5+KB (Pirtoaca)	85.4	11B
Flan-T5 (Chung et al., 2022)	86.5	3B
UnifiedQA (Khashabi et al., 2020)	87.2	11B
RumiFlanT5 (Yao et al., 2023)	87.3	3B
GenMC (Huang et al., 2022)	89.8	11B
LLM+RAG		
ReFeed	87.1	3B
FLARE	88.6	3B
Self-RAG	89.3	3B
Step-Back	89.0	3B
IEKR	**92.1**	3B

Table 2: Results on OpenbookQA dataset compared with LM+GNN and LLM+RAG based methods. 'Para" denotes the parameter number in the model. RumiFlanT5 are trained by using FlanT5-3B to replace RumiDeBERTa (Yao et al., 2023) as the backbone for fair comparison.

6 Analysis

In this part, we conduct ablation studies to evaluate different modules of our approach. Then we generalize our method to other QA tasks and compare our methods with different RAG based methods with the same backbone. Finally, we conduct experiments with different numbers of external knowledge sentences m.

6.1 Ablation Study

In our approach, there are 2 modules: Internal Knowledge Reflection, External Knowledge Retrieval. We successively evaluate the importance of different modules by removing the respect module.

Does internal knowledge reflection matter? In this ablation, we remove the process of digging into the LLM about the internal knowledge about the entities in query. We directly utilize the query to retrieve relevant external knowledge as inputs into LLM $\mathcal{M}$ along with the query to derive the answer. In Table 4, when removing the internal knowledge reflection, our model drops by 2.3 accuracy on CommonsenseQA, 3.4 accuracy on OpenbookQA, and 3.3 accuracy on MedQA. It demonstrates that without the internal knowledge within LLM, the retriever struggles to retrieve enough needed knowledge from external KB only with the query. However, compared with direct finetuning, this ablation model outperforms by 1.1, 2.2, and 0.9 accuracy on CommonsenseQA, OpenbookQA and MedQA, which shows our retriever still derives some useful external knowledge from KB $\mathcal{G}$ for the QA task.

Does external knowledge retrieval matter? In this ablation, we remove the process of retrieving knowledge from external KB and alleviating hallucination with verifier. We prompt the model to derive the internal knowledge about the query entity and input the internal knowledge with query to the LLM for the answer. In Table 4, when removing the external knowledge retrieval, our model drops by 1.5 accuracy on CommonsenseQA, 2.3 accuracy on OpenbookQA, and 2.7 accuracy on MedQA. It

demonstrates that the knowledge derived from internal reflection does not contain the enough information to answer the question. The model still needs to retrieve from external KB to complement the internal knowledge for knowledge-intensive QA. However, compared with direct finetuning, this ablation model outperforms by 1.9, 3.3, and 1.5 accuracy on CommonsenseQA, OpenbookQA and MedQA, which demonstrates the effectiveness of prompting LLM to reflect on internal knowledge.

6.2 Comparing with RAG methods

In this part, we compare our method with existing competitive RAG based methods ReFeed (Yu et al., 2023) and FLARE (Jiang et al., 2023) with the same backbone.

In Table **??**, it shows our method significantly outperforms the two competitive RAG baselines by over 2.2 accuracy on different datasets. Our method does not need the LLM to have the ability to give an initial answer or ask follow-up query. We focus on the concrete factual knowledge within LLM, which provides more new and valuable information for external retrieval. Compared with FLARE, the follow-up queries generated by smaller LLM like LLama other than ChatGPT do not provide as much valuable information as our method for external retrieval. Moreover, our method does not need multi-time retrieval and only retrieves once based on the concrete internal knowledge within LLM but derives significant improvement. Considering the large size of external KB, We reduce the computation cost and improve the performance by utilizing internal knowledge for external retrieval in one time. Compared with ReFeed, the concrete internal knowledge in LLM provides more valuable information for external retrieval than the initial answer of LLM. It demonstrates the significant effectiveness and efficiency of our method over other RAG methods. We highlighted the concrete factual knowledge within LLM instead of the initial answer or follow-up query of LLM in FLARE and ReFeed. Considering the setting of multi-choice QA, the choices has been included in query and initial answer of LLM will not provide much new knowledge beyond the content of query. However, our method prompts the LLM to provide concrete internal knowledge about the linguistic component of query. For example, we ask the LLM "tell me something about heat travel" and the LLM provides the internal knowledge "Heat travel through a thermal conductor" instead of "the answer is B" in ReFeed. To our best knowledge, we are the first to introduce concrete internal knowledge within LLM for external retrieval.

Methods	Acc.	Params
LM+GNN		
SapBERT (Liu et al., 2020)	37.2	-
QA-GNN (Yasunaga et al., 2021)	38.0	-
GREASELM (Zhang et al., 2022)	38.5	-
GrapeQA (Taunk et al., 2023)	39.5	-
LLM		
GPT-Neo (Black et al., 2022)	33.3	2.7B
LLama2 (Touvron et al., 2023)	46.7	7B
RumiLLama2 (Yao et al., 2023)	47.4	7B
LLM+RAG		
ReFeed	47.1	7B
FLARE	48.2	7B
Self-RAG	48.0	7B
Step-Back	47.6	7B
IEKR	**50.9**	7B

Table 3: Results on MedQA dataset compared with LM+GNN and LLM+RAG based methods. "Params" denotes the parameter number in the model. RumiLLama2 are trained by using LLama2-7B to replace RumiDeBERTa (Yao et al., 2023) as the backbone for fair comparison. We conduct experiments with 5 different random seeds and the p-value is less than 0.001.

On the other hand, we design the module of verifier shared by different datasets to modify the internal knowledge based on the retrieved external knowledge and combine the two sources of knowledge to answer the question. In FLARE and ReFeed, the internal knowledge contains the initial answer or follow-up query and may introduce distracting contents especially for smaller LLM like Llama-7b.

Retrieval from external KB can bring much computation cost because of the vast amount of knowledge. Our method does not need multi-time retrieval like FLARE and only retrieves once based on the concrete internal knowledge within LLM to derive significant improvement.

6.3 Generalization to Other QA Tasks

In this section, we apply our methods to the widely-recognized open-domain question answering (QA) dataset, 2WikiMultihopQA (**?**). Our approach follows the framework established by FLARE, utilizing the Llama-7b model as the foundational backbone, to showcase the generalization capabilities

Methods	CSQA	OBQA	MedQA
- internal	90.4	88.7	47.6
- external	90.2	89.8	48.2
- self-conclude	91.8	91.0	48.8
= self-critic	92.1	90.8	48.6
IEKR	**93.7**	**92.1**	**50.9**

Table 4: Ablation results on CommonsenseQA IHtest set, OpenbookQA test set, and MedQA test set. "CSQA" denotes CommonsenseQA, "OBQA" denotes OpenbookQA. "-internal" denotes removing the internal knowledge relection; "- external" denotes removing external knowledge retrieval; "Backbone" denotes directly finetuning the backbone model to generate the answer, i.e., FlanT5-3B for CSQA and OBQA.

Number	CSQA	OBQA	MedQA
10	92.9	91.0	49.5
30	93.2	91.7	50.3
50	**93.7**	92.1	**50.9**
100	93.1	**92.2**	50.4

Table 5: Results with different number of external knowledge sentences. We use accuracy on test set as evaluation.

of our methodology. Unlike multiple-choice question answering (MCQA), where the objective is to select the correct answer from a given set of options, open-domain QA requires generating free-form text responses to queries. To evaluate the performance of our method, we employ Exact Match (EM) and F1 score as the primary metrics.

As illustrated in Table 6, our method surpasses the performance of FLARE, achieving an improvement of 1.4 in EM and 1.9 in F1 score. Notably, while FLARE necessitates multiple retrievals of external knowledge to formulate an answer, our method accomplishes this with a single retrieval step. This significant difference underscores the efficiency and effectiveness of our approach when applied to the open-domain QA dataset. The ability to reduce retrieval frequency without compromising accuracy not only highlights the robustness of our method but also suggests potential for enhanced scalability and practicality in real-world applications.

6.4 Different Retrieval Number

In this section, we examine the impact of varying the number of external knowledge sentences, denoted as m, on our method's performance. For our primary experiments, we set m to 50, and we conducted additional experiments varying m from 10 to 100. As shown in Table 5, we observed that generally, retrieving a greater number of knowledge sentences from the external knowledge base (KB) enhances the performance of our method. The inclusion of more external knowledge provides the model with a richer set of factual information pertinent to the query, thereby improving the verifier's ability to mitigate hallucinations when revising internal knowledge.

However, an interesting phenomenon occurs when m is set to 100: the Integrated External Knowledge Retriever (IEKR) exhibits a slight decline in performance. This drop can be attributed to the longer external knowledge context, which includes some relevant but unnecessary sentences. These extraneous sentences can distract the model, complicating its reasoning process and ultimately impairing its ability to derive accurate answers. Therefore, while increasing the amount of retrieved external knowledge generally benefits performance, there is a threshold beyond which the inclusion of superfluous information can become counterproductive. This finding underscores the importance of optimizing the balance between the quantity and relevance of external knowledge in enhancing the model's reasoning capabilities.

Model	EM	F1
ReFeed	61.5	66.2
FLARE	62.0	66.7
Self-RAG	61.7	65.8
Step-Back	62.1	65.9
Ours	63.4	68.6

Table 6: Evaluation on 2WikiMultihopQA dataset with Llama-7b as the backbone. We utilize Exact match and F1 as the evaluation metrics.

7 Conclusion

In this work, we propose the internal and external knowledge interactively refinement framework, where the internal knowledge within LLM are utilized to retrieve needed knowledge in external KB, and the external knowledge retrieved are incorporated into revising the internal knowledge. We demonstrate our effectiveness on 3 benchmark datasets in knowledge-intensive QA with different LLMs across different domains.

Limitations

We propose the internal and external knowledge interactively refinement framework, and demonstrate our effiveness on 3 benchmark datasets in knowledge-intensive QA with different LLMs across different domains. The sizes of LLMs we use range from 3B to 7B, and we will conduct experiments with LLM larger than 7B in future research.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

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

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. *arXiv preprint arXiv:2005.00646*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zixian Huang, Ao Wu, Jiaying Zhou, Yu Gu, Yue Zhao, and Gong Cheng. 2022. Clues before answers: Generation-enhanced multiple-choice qa. *arXiv preprint arXiv:2205.00274*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Yanyang Li, Jianqiao Zhao, Michael R Lyu, and Liwei Wang. 2022. Eliciting knowledge from large pre-trained models for unsupervised knowledge-grounded conversation. *arXiv preprint arXiv:2211.01587*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

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.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

George Sebastian Pirtoaca. Ai2 leaderboard. *URL https://leaderboard. allenai. org/open_book_qa/submission/brhieieqaupc4cnddfg0*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Dhaval Taunk, Lakshya Khanna, Siri Venkata Pavan Kumar Kandru, Vasudeva Varma, Charu Sharma, and Makarand Tapaswi. 2023. Grapeqa: Graph augmentation and pruning to enhance question-answering. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1138–1144.

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

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022. Pinto: Faithful language reasoning using prompt-generated rationales. *arXiv preprint arXiv:2211.01562*.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. *arXiv preprint arXiv:2005.00691*.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. 2019. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7208–7215.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Weiwen Xu, Huihui Zhang, Deng Cai, and Wai Lam. 2021a. Dynamic semantic graph construction and reasoning for explainable multi-hop science question answering. *arXiv preprint arXiv:2105.11776*.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2021b. Human parity on commonsenseqa: Augmenting self-attention with external attention. *arXiv preprint arXiv:2112.03254*.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021c. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1201–1207.

Jun Yan, Mrigank Raman, Aaron Chan, Tianyu Zhang, Ryan Rossi, Handong Zhao, Sungchul Kim, Nedim Lipka, and Xiang Ren. 2020. Learning contextualized knowledge structures for commonsense reasoning. *arXiv preprint arXiv:2010.12873*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yunzhi Yao, Peng Wang, Shengyu Mao, Chuanqi Tan, Fei Huang, Huajun Chen, and Ningyu Zhang. 2023. Knowledge rumination for pre-trained language models. *arXiv preprint arXiv:2305.08732*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

A Appendix

Dataset	Example
CommonsenseQA	A weasel has a thin body and short legs to easier burrow after prey in a what? (A) tree (B) mulberry bush (C) chicken coop (D) viking ship (E) rabbit warren
OpenbookQA	Which of these would let the most heat travel through? (A) a new pair of jeans (B) a steel spoon in a cafeteria (C) a cotton candy at a store (D) a calvin klein cotton hat
MedQA-USMLE	A 57 -year-old man presents to his primary care physician with a 2-month history of right upper and lower extremity weakness. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had increasing difficulty with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of his family have had musculoskeletal problems. His right upper extremity shows forearm atrophy and depressed reflexes while his right lower extremity is hypertonic with a positive Babinski sign. Which of the following is most likely associated with the cause of this patients symptoms? (A) HLA-B8 haplotype (B) HLA-DR2 haplotype (C) Mutation in SOD1 (D) Mutation in SMN1

Table 7: Examples of the Knowledge-intensive QA task for each of the datasets evaluated in this work.