Evaluating the External and Parametric Knowledge Fusion of Large Language Models

Hao Zhang[*], **Yuyang Zhang**[*], **Xiaoguang Li**, **Wenxuan Shi**, **Haonan Xu**, **Huanshuo Liu**
Yasheng Wang, **Lifeng Shang**, **Qun Liu**, **Yong Liu**, **Ruiming Tang**
Noah's Ark Lab, Huawei Technologies Co., Ltd
{zhang.hao3,zhangyuyang4}@huawei.com

Abstract

Integrating external knowledge into large language models (LLMs) presents a promising solution to overcome the limitations imposed by their antiquated and static parametric memory. Prior studies, however, have tended to over-reliance on external knowledge, underestimating the valuable contributions of an LLMs' intrinsic parametric knowledge. The efficacy of LLMs in blending external and parametric knowledge remains largely unexplored, especially in cases where external knowledge is incomplete and necessitates supplementation by their parametric knowledge. We propose to deconstruct knowledge fusion into four distinct scenarios, offering the first thorough investigation of LLM behavior across each. We develop a systematic pipeline for data construction and knowledge infusion to simulate these fusion scenarios, facilitating a series of controlled experiments. Our investigation reveals that enhancing parametric knowledge within LLMs can significantly bolster their capability for knowledge integration. Nonetheless, we identify persistent challenges in memorizing and eliciting parametric knowledge, and determining parametric knowledge boundaries. Our findings aim to steer future explorations on harmonizing external and parametric knowledge within LLMs.

1 Introduction

Here is a potential to do analysis work

Parametric knowledge acquired by large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023; Anil et al., 2023; Du et al., 2022) during pre-training inevitably becomes outdated over time. Integrating additional contents into LLM inputs has emerged as an effective strategy to mitigate such issue (Lewis et al., 2020; Nakano et al., 2021; Gao et al., 2023). By incorporating external knowledge either into the input context (Ram et al., 2023; Izacard et al., 2022) or through intermediary layers (Borgeaud et al., 2022; Wu et al., 2022), LLMs are endowed with more current information, expanding their knowledge boundary and reducing the instances of hallucinations and factual errors.

Many retrieval (Lewis et al., 2020; Asai et al., 2023; Izacard et al., 2022) or tool (Shen et al., 2023; Qin et al., 2024; Schick et al., 2023) augmented methods predominantly rely on external evidence and often overlooking the rich knowledge stored within LLMs. Yet, the external evidence, inevitably, could be incomplete and noisy. While some approaches propose to refine the external evidence and post-calibrate the outputs by tapping into LLMs' parametric knowledge (Meng et al., 2022; Zhang et al., 2024), the full potential of merging external with parametric knowledge remains unexplored. This paper aims to delve into *how LLMs perform external and parametric knowledge fusion across various conditions*, especially when **LLMs encounter incomplete or irrelevant external knowledge**. A thorough understanding of this is crucial for a broader application of knowledge-augmented LLMs. Not only does this relate to the LLMs' parametric memory elicitation (Xie et al., 2024; Qian et al.,

[*]The first two authors contributed equally.

Preprint. Work in progress.

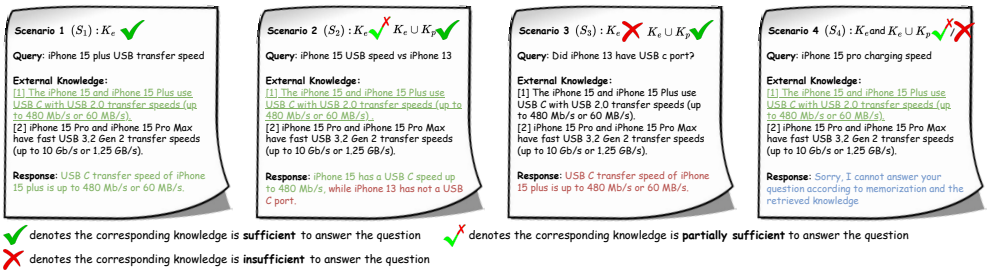

Figure 1: An illustration of four parametric and external knowledge fusion scenarios in LLMs.

2023; Wang et al., 2023b,c), but it is also associated with the knowledge boundary perception of LLMs (Ren et al., 2023b; Zhang et al., 2023b; Yin et al., 2023b).

To elucidate the dynamics of LLMs in integrating external (K_e) and parametric (K_p) knowledge[2], we define four distinct scenarios reflecting the interplay between K_e and K_p (depicted in Figure 1). The scenarios are as follows: (1) S_1 indicates that K_e alone is sufficient to answer a query, independent of K_p's contribution; (2) S_2 suggests that K_e provides partial information, requiring K_p to fill the gaps for a complete answer; (3) S_3 identifies situations where K_e offers no useful information, and the answer depends solely on K_p; (4) S_4 describes cases where neither K_e nor K_p adequately address a query, making it theoretically unanswerable. Prior studies (Yoran et al., 2023; Chen et al., 2023) often presume situations where the availability of external knowledge (K_e) and K_p is non-contributory, simplifying the knowledge fusion process to scenarios S_1 and S_4 and neglecting intermediate cases. The real challenge emerges when K_e is sub-optimal, necessitating a nuanced integration of K_e and K_p for a cooperative response, especially in scenarios S_2 and S_3. However, the model-specific nature of LLMs' K_p significantly complicates the precise delineation of knowledge boundaries and access to parametric knowledge. This complexity impedes a thorough and impartial evaluation of LLMs' capabilities in knowledge fusion.

To mitigate the challenges associated with acquiring parametric knowledge by LLMs, we propose a systematic pipeline for data construction and knowledge infusion. Specifically, we first collect the latest data from the electronic product domain and divide it into two parts: one for enhancing LLMs' parametric knowledge (K_p) through continued training, and the other as external knowledge (K_e). We also craft a set of questions based on the data to emulate the four scenarios: queries that solely depend on K_e (S_1), queries requiring integration of K_e and K_p (S_2), queries dependent only on K_p (S_3), and unanswerable queries (S_4). For each scenario, we provide relevant evidence and introduce additional distractors to mimic real-world conditions. Overall, this aims to standardize the parametric knowledge within different LLMs, facilitating equitable and model-independent evaluations.

active learning

We first inject new knowledge into LLMs through continued training and supervised fine-tuning, subsequently evaluating their knowledge retention. Then, we design a series of experiments to reveal the behaviors of LLMs in knowledge fusion. Despite the performance gains by integrating external and parametric knowledge, the results indicate that: (1) LLMs show deficiencies in recognizing domain knowledge, significantly influenced by their capacity to retain knowledge. (2) There are persistent challenges in memorizing and eliciting parametric knowledge and determining parametric knowledge boundaries for effective knowledge fusion. Our contributions are as follows:

① domain-specific ② boundary

- We review knowledge fusion in LLMs, defining four distinct scenarios reflecting the interplay between external and parametric knowledge fusion for thorough evaluation.
- To mitigate the challenges associated with acquiring parametric knowledge by LLMs, we propose a systematic pipeline for data construction and knowledge infusion to facilitate knowledge fusion exploration.
- Through extensive experiments on various LLM backbones, we identify persistent challenges in memorizing and eliciting parametric knowledge and determining parametric knowledge boundaries. These challenges impair the effectiveness of knowledge fusion.

[2]For simplicity throughout this paper, we use K_e and K_p to symbolize the retrieved external knowledge and the LLMs' parametric knowledge, respectively.

2 Related Work

Retrieval-augmented LLMs (RA-LLM). RA-LLM, including tools, are considered essential for linking LLMs with external knowledge sources (Lewis et al., 2020; Qin et al., 2024; Mialon et al., 2023; Gao et al., 2023), which makes LLMs more viable for practical applications. The prevalent methods either augment external evidence via in-context learning paradigm (Lazaridou et al., 2022; He et al., 2022; Izacard et al., 2022) or adopt external evidence to post-calibrate the generations (Meng et al., 2022; Li et al., 2024; Yan et al., 2024). Some work also suggests fine-tuning LLMs to enhance the utilization of external knowledge and optimize the retrieval strategy (Lewis et al., 2020; Borgeaud et al., 2022; Asai et al., 2023; Lin et al., 2023). These approaches mainly rely on external knowledge while overlooking the knowledge stored within LLMs, which may lead to undesirable results due to the biased and noisy external information (Mallen et al., 2022; Yoran et al., 2023; Liu et al., 2023b).

Parametric Knowledge in LLMs. After pre-training, LLMs have internalized massive knowledge into their parameters, *i.e.*, parametric knowledge (Petroni et al., 2019; Geva et al., 2021b; Hu et al., 2023; Gueta et al., 2023). However, recent studies indicate that effectively leveraging LLMs' parametric knowledge is challenging (Wang et al., 2023a; Allen-Zhu and Li, 2023a). That is, although LLMs can memorize extensive knowledge, it does not guarantee their ability to adeptly elicit and manipulate it for subsequent tasks (Berglund et al., 2023; Liu et al., 2023a; Wang et al., 2023d; Allen-Zhu and Li, 2023b). Some work also observes that compared to directly augmenting knowledge into inputs, LLMs struggle to accurately memorize knowledge into parameters (Kandpal et al., 2023; Ovadia et al., 2023). Besides, several studies explore the self-calibration (Rajpurkar et al., 2018; Kadavath et al., 2022; Yin et al., 2023a) and knowledge boundary detection (Ren et al., 2023a) in LLMs, benefit for improving confidence and interpretability in their use of parametric knowledge, thus reducing hallucinations. Similarly, our objective is to investigate the capacity of LLMs for knowledge memorization and utilization in the knowledge fusion process, along with their self-calibration ability.

Knowledge Fusion of LLMs. To perform the fusion of external and parametric knowledge, Jiang et al. (2023) propose dynamically assessing the confidence level of model generation and intervening with retrieval at low confidence. Wang et al. (2023c) elicit LLMs' ability to recognize their self-knowledge and achieve better knowledge integration. Some studies explore the knowledge conflict issues when integrating the external and parametric knowledge (Li et al., 2022; Pan et al., 2021; Mallen et al., 2022; Zhang et al., 2023a; Xie et al., 2023). However, these approaches mainly optimize knowledge fusion to enhance the subsequent tasks, such as open-domain QA (Kwiatkowski et al., 2019; Yang et al., 2018; Geva et al., 2021a), lacking a comprehensive evaluation of LLMs' behaviors in knowledge fusion. In contrast, we focus on the investigation of external and parametric knowledge fusion, including the systematic task definition, data construction pipeline, and thorough experiments.

3 Task Definition

In practical applications, the external evidence obtained through retrieval or tools may be noisy, incomplete, or irrelevant Yoran et al. (2023); Liu et al. (2023b). This leads to the necessity of thoroughly considering various conditions when evaluating the external and parametric knowledge fusion. Therefore, we define four distinct scenarios capturing the diverse interactions between external and parametric knowledge of LLMs, aiming to encompass all potential circumstances as comprehensively as possible. Given external knowledge K_e and parametric knowledge K_p, the defined scenarios are: (1) S_1 indicates K_e alone is sufficient to answer a query, independent of K_p's contribution; (2) S_2 denotes K_e carries partial information, requiring K_p to fill the gaps for a complete answer; (3) S_3 identifies situations where K_e offers no useful information, and the answer depends solely on K_p; (4) S_4 describes cases where neither K_e nor K_p adequately address a query, making it theoretically unanswerable.

Suppose K_p has been injected into the LLM. Formally, given a question q_{S_i} and the corresponding external evidence K_e^i, where $i \in \{1, 2, 3, 4\}$, the response $\hat{a}_{S_i}$ of an LLM is generated as:

$$\hat{a}_{S_i} = \texttt{LLM}_{(K_p)}([q_{S_i}; K_e^i; \texttt{inst}]), \tag{1}$$

where $\texttt{LLM}_{(K_p)}$ denotes the LLM already encoded the K_p into its parameters, `inst` represents the task-specific instructions. Ideally, for S_1, K_e^1 contains ground-truth evidence, where LLM can solely

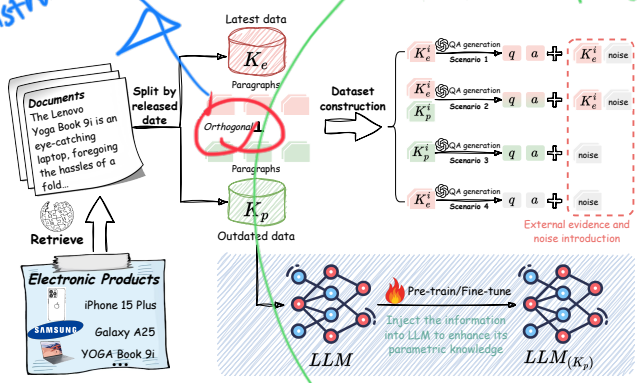

Figure 2: The overview of dataset construction. We first retrieve documents of electronics from websites. The documents are split into two portions based on their released date, and decomposed into paragraphs. The QA pairs for each scenario are generated via prompting LLMs, and the corresponding external evidence and noise are added as support sources. The outdated data is injected into LLMs through pre-training or fine-tuning.

depend on K_e^1 to accurately answer q_{s_1}, that is, $\hat{a}_{S_1}$ is aligned with ground-truth a_{S_1}. For S_2, K_e^2 holds only partial information relevant to q_{S_2}, LLM also requires to elicit its corresponding K_p to derive an accurate $\hat{a}_{S_2}$ for q_{s_2}. For S_3, K_e^3 is devoid of relevant information and is solely comprised of distractions, LLMs must eliminate these distractions and elicit its K_p to reach the correct $\hat{a}_{S_3}$. In S_4, where K_e^4 consists solely of distractors and LLM lacks relevant K_p, it should opt to refrain from responding, implying that $\hat{a}_{S_4}$ should incorporate a refusal to answer q_{S_4}. Following the criteria outlined, we construct datasets, fine-tune different LLMs, and conduct a detailed evaluation of their ability to integrate external and parametric knowledge in these scenarios.

4 Dataset Construction

Although LLMs encode massive knowledge through large-scale pre-training, the parametric knowledge of different LLMs exhibits notable variations due to discrepancies in training corpora, model scale, and forgetting issue (Wang et al., 2023a; Luo et al., 2023). Thus, it is challenging and almost infeasible to directly elicit the parametric knowledge of various LLMs (Qian et al., 2023) and interact with external knowledge to conduct a fair and comprehensive evaluation.

In this work, we focus on the assessment of knowledge fusion under a standard RAG setting. Facing difficulties in acquiring the parametric knowledge, we instead collect data to enrich LLMs' knowledge, enabling controlled and quantifiable evaluation of knowledge fusion. We split the data into two partitions, one part serving as external knowledge (K_e), and the other part integrated into the LLMs as parametric knowledge (K_p) through training. In this way, we eliminate the inconsistencies in K_p among different LLMs. Leveraging the collected information, we further employ LLM to generate relevant question-answer pairs, forming a standard QA dataset for subsequent training and evaluation.

Data Source Preparation. LLMs trained on cut-dated data are rarely exposed to domain-specific knowledge and have not yet encountered the latest information. Thus, the most recent and high-quality data sources are essential for building a viable dataset in our setting. Considering the swift evolution, variety, and annual surge of new products in the electronics domain, it emerges as a suitable source to measure knowledge fusion. Thus, we collect the data in the electronics domain spanning the preceding four years and utilize product introductory documents with detailed specifications serving as the primary source. Specifically, we collect over 500 mobile phone names from websites and execute online searches to collate multiple search results for each product. Then, document filtration is applied through empirical rules and manual review, preserving documents with unique product introductions. Since some documents are too lengthy, we dissect them at the granularity of paragraphs and sentences to extract varied relevant information for each product. In general, we filtered out $1,700$ paragraphs from $5,000$ paragraphs in $1,500$ documents, where 900 paragraphs are used to construct external knowledge, while 800 paragraphs are trained in LLMs as parametric knowledge.

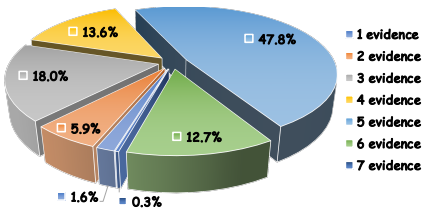

Figure 3: The distribution of the number of the associated evidence per QA sample.

Dataset Construction. The overview of dataset construction pipeline is shown in Figure 2. To simulate the external and parametric knowledge fusion scenarios, we divide the collected data into *latest* and *outdated* data according to their released date[3]. We highlight that the outdated data may be learned by LLMs, while the latest data is less likely seen by LLMs due to cut-dated pre-training. We retain the *latest data as external knowledge*, while the *outdated data as parametric knowledge*, which is injected into LLM to enhance its parametric memory via fine-tuning. The evaluation of the LLMs' competence in fusing external and parametric knowledge focuses on its discriminating usage of external and parametric knowledge, as well as the correctness of the information utilized. To more clearly assess LLM capabilities, we use the partitioned latest and outdated data to develop respective QA evaluation datasets with sophisticated designed instructions. Specifically, we generate the QA pairs for four knowledge fusion scenarios as follows:

- **Scenario 1** (S_1): We randomly select one or two snippets from the latest data to generate a QA pair with an LLM, and then the unrelated snippets are added as noise to the chosen snippets, creating the candidate knowledge for the generated QA pair.
- **Scenario 2** (S_2): We randomly select one snippet from both the latest and outdated data and generate a QA pair based on the two snippets. Unrelated snippets are added as noise to form the candidate knowledge for the QA pair.
- **Scenario 3** (S_3): The QA pair is generated solely based on the chosen snippets from outdated data, and noisy snippets are added to form candidate knowledge.
- **Scenario 4** (S_4): We randomly select one or two snippets from the latest data to create a QA pair, then discard the snippets and choose unrelated noise snippets as candidate knowledge for the QA pair, ensuring the unanswerability of the generated question.

To better align with real-world application settings, we adopt two noise introduction approaches to increase the challenge for LLMs in leveraging external knowledge. The first approach introduces noise snippets describing identical attributes across different electronic products, whereas the second approach presents noise snippets describing disparate attributes of a single electronic product. Moreover, to guarantee data quality, we further employ LLM evaluation coupled with manual review for data cleansing and filtering, yielding 210, 580, 200, and 140 samples per scenario.

Dataset Analysis. We first assess the distribution of external evidence associated with each entity. Entities with three pieces of evidence are most prevalent, constituting 19.5% of the sample, followed by entities with a single piece of evidence, accounting for 17.5% of the total. We also examine the distribution of associated evidence per sample, depicted in Figure 3. Notably, 47.8% of samples contained five pieces of evidence, the highest proportion, while samples with three pieces of evidence comprised the second-highest percentage at 18%. Subsequently, we analyze the distribution of evidence lengths across the dataset, finding that quotes ranging from 500 to 600 characters represented the majority, totaling 78.6%. The overview of dataset partitioning is shown in Table 1. The dataset comprises training, validation, and test sets with 630, 300, and 300 samples, respectively. The table also details the distribution of data across scenarios S_1, S_2, S_3, and S_4 within each subset.

Data Split	S_1	S_2	S_3	S_4	Total
train	100	390	90	50	630
dev	50	150	50	50	300
test	60	140	60	40	300

Table 1: The statistics of the dataset in each scenario.

5 Experiment Setup

Backbone Model. We select the open-source ChatGLM3-6B (Du et al., 2022) and Qwen-7B (Bai et al., 2023), and the black-box GPT-4 (OpenAI, 2023) as the backbones. These models are selected for their robust language understanding and instruction-following capabilities, which align well with our experiment design. Furthermore, the open-source LLMs enable flexible adaptation of model configurations and the analysis of their internal behaviors.

[3]The latest data is orthogonal to outdated data, *i.e.*, no information overlap between them. Besides, the latest data refers to factual information about electronic products occurring after 2023-06-01, unseen by existing LLMs.

Table 2: Knowledge infusion results of ChatGLM Du et al. (2022) and Qwen Bai et al. (2023).

Model	Accuracy (%)	Coverage		
		Complete (%)	Partial (%)	Uncover (%)
ChatGLM	13.3	3.3	25.0	71.7
$\text{ChatGLM}_{\text{CT}}$	38.3	18.3	36.7	45.0
Qwen	15.0	5.0	21.7	73.3
Qwen_{CT}	43.3	20.0	43.3	36.7

Parametric Knowledge Infusion. For the selected LLMs, the outdated data portion needs to be injected into them via continued training or fine-tuning[4]. Since LLMs predominantly acquire knowledge in the pre-training phase, we also employ the same strategy for continued training. However, Allen-Zhu and Li (2023a) indicates that "memorization of knowledge" in language models merely means the model can fit the exact training data but does not imply it can extract the knowledge flexibly from data after training. To enhance knowledge memorization, we further adopt the data rewriting strategy suggested in Allen-Zhu and Li (2023a) to conduct data augmentation. Specifically, we use GPT-4 (OpenAI, 2023) to paraphrase the snippets in the outdated data portion and generate eight QA pairs related to that snippet as the supplementary data. The synthetic data is merged with the original data to train the backbones.

Evaluation Metrics. We employ accuracy (R_{acc}) and information coverage (R_{cover}) as evaluation metrics to access the knowledge fusion capabilities of LLMs. Accuracy assesses if LLM responses accurately address the question and align with both external and parametric knowledge sources. Responses are deemed correct if consistent with these sources and incorrect if they include irrelevant content or deviate from the information provided. Information coverage refers to the degree to which LLMs encapsulate the core content of the reference. This coverage is classified into three categories: complete, partial, and no inclusion. Let K_{gen} denote the knowledge contained in generations, K_{gold} indicate the knowledge contained in the ground-truth answer, and K_{ref} represents the given external and parametric knowledge[5]. The R_{acc} and R_{cover} are computed as follows:

$$R_{\text{acc}} = \begin{cases} 1 & \text{if } K_{\text{gen}} \subseteq K_{\text{ref}}, \\ 0 & \text{otherwise} \end{cases} \quad (2) \qquad R_{\text{cover}} = \begin{cases} \text{Complete} & \text{if } K_{\text{gold}} \subseteq K_{\text{gen}} \\ \text{Partial} & \text{if } K_{\text{gold}} \cap K_{\text{gen}}. \\ \text{Uncover} & \text{otherwise} \end{cases} \quad (3)$$

6 Experiment Results and Analysis

In this section, we conduct comprehensive experiments and in-depth analysis to investigate the knowledge fusion behaviors of various backbones. We use $\text{ChatGLM}_{\text{CT}}$ to represent ChatGLM that continues trained on K_p, and $\text{ChatGLM}_{\text{CT\&SFT}}$ denotes ChatGLM that continues trained on K_p and further SFT on our train set. Similar to Qwen_{CT} and $\text{Qwen}_{\text{CT\&SFT}}$.

6.1 The Performance of Knowledge Infusion

To investigate the effectiveness of knowledge infusion by LLMs at continued training process, we conduct experiments using two models: ChatGLM3-6B (*abbr.* ChatGLM) and Qwen-7B (*abbr.* Qwen). These models are trained using the designated parametric knowledge partition, K_p. The evaluation involved querying the models with questions specifically related to K_p to assess how well the models retained the trained knowledge. It is important to note that this evaluation is similar to scenario S_3, absent the inclusion of external knowledge distractors. We adopt the question-answer (QA) pairs from S_3 by excluding the associated external evidence to serve as our evaluation dataset.

The results, summarized in Table 2, reveal that before continued training, both models demonstrated notably low accuracy rates: 13.3% for ChatGLM and 15.0% for Qwen, suggesting these models

[4]Given the inaccessibility of GPT-4's weights, we assume it already memorizes outdated data and only conducts inference by providing external knowledge snippets as evidence. `GPT-4-0613` is used in all experiments.

[5]When evaluating LLMs' competence in S_4, we only measure their capability to decline to respond correctly.

Table 3: The overall performance of different LLMs under four scenarios. "Direct" represents directly prompting LLM to answer the questions by giving the corresponding external knowledge without continued training and supervised fine-tuning; "SFT" denotes the supervised fine-tuning on the train set of our constructed question-answering dataset; and "CT" means continuing training on the K_p data partition to inject the knowledge into the LLM; "Easy" denotes the supervised fine-tuning on the train set as well as providing the supporting snippets during inference.

Scenario	Metric	GPT-4	ChatGLM				Qwen	
			Direct	SFT	CT&SFT	Easy	SFT	CT&SFT
S_1	R_{acc} (%)	81.7	63.3	68.3	61.7	72.7	62.9	63.3
	Complete (%)	80.0	38.3	38.3	33.3	43.3	31.7	30.0
	Partial (%)	11.7	40.0	48.3	55.0	35.0	56.7	53.3
	Uncover (%)	8.3	21.7	13.4	11.7	21.7	11.6	16.7
S_2	R_{acc} (%)	35.7	39.3	52.1	53.6	72.1	49.3	57.1
	Complete (%)	12.9	9.3	7.1	20.0	42.1	10.0	22.1
	Partial (%)	40.0	56.4	76.4	69.3	40.0	71.5	61.4
	Uncover (%)	47.1	34.3	16.5	10.7	17.9	18.5	16.5
S_3	R_{acc} (%)	8.3	10.0	16.7	35.0	78.3	20.0	33.3
	Complete (%)	3.3	1.7	3.3	16.7	55.0	3.3	20.0
	Partial (%)	11.7	23.3	40.5	45.0	30.0	48.3	41.7
	Uncover (%)	85.0	75.0	56.2	38.3	15.0	48.4	38.3
S_4	R_{acc} (%)	37.5	25.0	30.0	40.0	-	27.5	40.0

indeed have no such background knowledge. After continued training, there was a substantial enhancement in performance. Specifically, ChatGLM exhibits a 25% absolute improvement in accuracy, while Qwen shows a 28.3% absolute increase. This significant enhancement underscores the efficacy of continued training in injecting the knowledge into the models. Meanwhile, there is a notable enhancement in the model's ability to answer questions with complete and partial accuracy after continued training. Specifically, ChatGLM displays increases of 15% and 11.7% in complete and partial correct responses respectively, whereas Qwen showed improvements of 15% and 21.6%.

Ideally, knowledge infusion through continued training of an LLM should enable the model to retain all imparted knowledge, resulting in the QA accuracy nearing 100%. In practice, however, even though accuracy significantly improves over untrained models, it remains considerably lower than the optimal situation. This suggests substantial amounts of knowledge are either not retained or not accurately elicited by the LLM. We highlight two key factors attributed to this issue: (i) model capability and (ii) dataset diversity. For the model capability, recent studies (Allen-Zhu and Li, 2023a,b) highlight that LLM faces difficulties using its parametric knowledge, and processing such knowledge does not guarantee it to be elicited accurately. For the dataset diversity, the LLM simply memorizes the given knowledge, meaning it only fits the given contents and may not effectively utilize this knowledge. For instance, LLM is trained in a massive of documents during continued training and evaluated under the question-answering manner at test time, LLM may not effectively map the questions to the answers learned during training. Besides, altering the way questions are posed might prevent the LLM from providing correct answers, and the reversal curse (Berglund et al., 2024) is another example of such an issue. Thus, a straightforward solution is to diversify the given knowledge, such as constructing various types of QA pairs, paraphrasing the documents, etc., that training LLM to memorize the knowledge from different perspectives.

6.2 Main Results

In this section, we conduct a comprehensive evaluation of different LLMs over the four knowledge fusion scenarios. The results are summarized in Table 3. Note "Direct" mode denotes that we directly prompt LLM to answer the questions by giving the corresponding external knowledge without continued training or supervised fine-tuning; "SFT" mode represents that we supervised fine-tune our constructed train set; "CT&SFT" mode denotes that we continue training the LLM on K_p following by further supervised fine-tune on the constructed train set; and "Easy" mode means that we not only

SFT the LLM on the constructed train set but also provide the ground-truth snippets coupled with distractors during inference.

6.2.1 Knowledge Fusion Performance on S_1

Scenario 1, S_1, denotes that provided external knowledge, K_e, alone is sufficient to answer a question, independent of K_p's contribution. The S_1 results of the different models are summarized in Table 3. Observed that GPT-4 achieves the best performance among all models, which obtains 81.7% accuracy and 66.7% complete coverage. The higher accuracy usually leads to better "complete" and/or "partial" coverage. Compared to ChatGLM and Qwen, GPT-4 has a richer internal knowledge base and more powerful content comprehension capabilities. For ChatGLM, $\text{ChatGLM}_{\text{SFT}}$ is superior to $\text{ChatGLM}_{\text{Direct}}$, *i.e.,* vanilla ChatGLM without continued train or SFT, by 5% absolute improvements on accuracy. Notably, the continued training does not always contribute to the performance improvements in scenario 1 (S_1). For instance, $\text{ChatGLM}_{\text{SFT}}$ obtains 68.3% accuracy while $\text{ChatGLM}_{\text{CT\&SFT}}$ only achieves 61.7%. Qwen_{SFT} is comparable to the $\text{Qwen}_{\text{CT\&SFT}}$ with only 0.4% accuracy gap. We highlight it is because all the ground-truth evidence is provided by external knowledge in S_1, SFT helps LLM to learn how to follow the instructions and utilize the given knowledge to reach the correct responses. The knowledge provided by CT is useless in the S_1 scenario, and continued training may inevitably lead to capability degradation of LLM (Shi et al., 2024). Nevertheless, $\text{ChatGLM}_{\text{SFT}}$ is inferior to $\text{ChatGLM}_{\text{Easy}}$ with a distinct gap, *i.e.,* 68.3% versus 72.2%, which demonstrates that noisy external knowledge indeed affects LLMs adversely (Pan et al., 2024; Cuconasu et al., 2024). Notably, the noise in our dataset is carefully curated, being relevant yet useless for effective responses.

6.2.2 Knowledge Fusion Performance on S_2

Scenario 2, S_2, represents that K_e provides *partial* knowledge to answer a question, and it requires K_p to fill the gaps for a complete answer. As summarized in Table 3, the overall performance of different backbone models in S_2 is significantly inferior to that in S_1. For instance, although the test cases are different, the accuracy of GPT-4 drops from 81.7% to 35.7% and the complete coverage drops from 80.0% to 12.9%, which proves that our data partition, *i.e.,* K_e and K_p, is reasonable, indicating that the knowledge in K_p is not covered by the off-the-shelf LLMs. Compared to $\text{ChatGLM}_{\text{Direct}}$, $\text{ChatGLM}_{\text{SFT}}$ achieve much better performance, 52.1% versus 39.3%. Since SFT does not inject K_p into LLM, both $\text{ChatGLM}_{\text{Direct}}$ and $\text{ChatGLM}_{\text{SFT}}$ suffer from low complete coverage. Comparing $\text{ChatGLM}_{\text{SFT}}$ with $\text{ChatGLM}_{\text{CT\&SFT}}$, the complete coverage of $\text{ChatGLM}_{\text{CT\&SFT}}$ is higher than $\text{ChatGLM}_{\text{SFT}}$ by a large margin, which indicates that CT indeed injects the K_p into the LLMs, the LLMs are capable of using the knowledge to answer the questions by considering both its parametric knowledge and the given external knowledge. A similar observation is held for Qwen model.

However, the accuracy of CT&SFT is slightly better than that of SFT for both ChatGLM and Qwen models. We emphasize that there are two aspects to this issue. One primary factor is the model's memory capacity, which determines the extent of knowledge retained during training. As discussed in Section 6.1, due to limitations in model capacity and dataset diversity, the model can accurately retain only a subset of the provided K_p. Another factor is that LLMs face difficulties using their parametric knowledge (Allen-Zhu and Li, 2023a,b) and accurate parametric and external knowledge fusion for question answering is challenging. According to case studies in S_2, we observe that the success rate of LLM's parametric knowledge elicitation is only around 60%. If we directly use all the ground-truth supporting snippets as external knowledge and feed them into LLMs, the LLMs' performance increases significantly (see "Easy" and "CT&SFT"), which further proves the deficiency of LLMs to utilize their parametric knowledge. Some work (Jeong et al., 2024; Ding et al., 2024) performs parametric and external knowledge fusion by first producing partial answers using parametric knowledge and then integrating the generated knowledge and external knowledge for final answer generation. In contrast, we directly prompt LLM to generate the final answer by considering its parametric knowledge and the given external knowledge.

6.2.3 Knowledge Fusion Performance on S_3

Note that scenario 3, S_3, simulates the situation that K_e offers no useful information and the correct answer depends solely on K_p. As reported in Table 3, without SFT or CT, all the evaluated backbone models fail in S_3. For instance, GPT-4 and $\text{ChatGLM}_{\text{Direct}}$ only obtain 8.3% and 10.0% accuracy, respectively. Compared to $\text{ChatGLM}_{\text{Direct}}$, $\text{ChatGLM}_{\text{SFT}}$ slightly improves the performance, since

SFT only teaches LLM to follow the instruction for answer generation, while it does not inject the new knowledge into the LLM. After injecting the K_p into LLM, ChatGLM$_{\text{CT\&SFT}}$ significantly outperforms ChatGLM$_{\text{SFT}}$ in accuracy by 18.3% absolute improvement. A similar result is observed for the Qwen model, which obtains 13.3% absolute gains. Despite the improvements achieved, their performance is still sub-optimal. For instance, ChatGLM$_{\text{Easy}}$ reached 78.3% accuracy in S_3, which 43.3% higher than ChatGLM$_{\text{CT\&SFT}}$. Similar to the observation in Section 6.2.2, the results indicate that CT cannot guarantee that LLM will fully retain all knowledge, and LLM itself faces difficulties in accurately eliciting parametric knowledge. Moreover, in the knowledge infusion experiment (ref. Section 6.1), we use the same QA pairs as S_3, but we ignore all the external distractors. Comparing the results between knowledge infusion (see Table 2) and S_3 knowledge fusion (see Table 3), we observe that the accuracies of both ChatGLM and Qwen in S_3 knowledge fusion are lower than that in knowledge infusion, which emphasizes that incorporating noisy external knowledge negatively impacts LLM performance, as it may cause overconfidence in plausible but incorrect information.

6.2.4 Knowledge Fusion Performance on S_4

Recall that scenario 4, S_4, describes cases where neither K_e nor K_p adequately address the questions, making those questions theoretically unanswerable. This scenario aims to evaluate the efficacy of LLMs to correctly provide a refusal response if they do not have the corresponding parametric knowledge and the external knowledge is unhelpful. As reported in Table 3, all the evaluated backbone models, including GPT-4, fail to trigger the refusal response under S_4. In general, these models tend to be overconfident in the provided plausible but incorrect external knowledge, yielding wrong answers. SFT shows positive impacts on performance improvement, where ChatGLM$_{\text{SFT}}$ is 5% higher than ChatGLM$_{\text{Direct}}$. Due to the presence of some refusal response samples in the data, SFT can guide the LLM on how to trigger and issue refusals to some extent. Comparing ChatGLM$_{\text{CT\&SFT}}$ with ChatGLM$_{\text{SFT}}$, CT further boosts the performance. We speculate that continued training with domain knowledge improves the LLM's field-specific understanding, enhancing its ability to discern whether the provided external knowledge and its parametric knowledge can effectively address a given question.

6.3 Findings and Challenges

The knowledge is fuzzy.

According to the in-depth analyses presented in Section 6.1 and 6.2, we conclude the observations and insights as follows:

- **Noise Robustness of LLM**: Noise and interference information from external knowledge negatively impact LLM performance (Chen et al., 2024), as evidenced across multiple LLMs in scenarios $S_1 \sim S_4$, leading to the generation of seemingly plausible but incorrect answers.
- **Impact of supervised fine-tuning**: Across all scenarios, $S_1 \sim S_4$, supervised fine-tuning (SFT) helps to improve the performance of LLMs. Despite SFT (almost) does not inject new knowledge into the LLM, it enhances the LLM's ability in instruction adherence, leading to more standardized outcomes (Allen-Zhu and Li, 2023a,b).
- **Impact of continued training**: when external knowledge is sufficient, *i.e.,* S_1, domain knowledge infusion via continued training yields negligible improvement, as the LLM can generate correct answers based solely on the provided information. Conversely, when external knowledge is inadequate, *i.e.,* $S_2 \sim S_4$, continued training is crucial and significantly enhances performance (Jiao et al., 2023; Naveed et al., 2024; Fujii et al., 2024), since LLM lacks the necessary domain knowledge and continued training can effectively alleviate the knowledge limitations of LLM.
- **The effect of knowledge infusion**: Although experiments on knowledge infusion and $S_2 \sim S_3$ demonstrate the effectiveness, performance gains of knowledge infusion remain limited. Due to constraints in model capacity and dataset diversity, LLMs can retain only a subset of the knowledge accurately via continued training (Moiseev et al., 2022; Arrotta et al., 2024). Additionally, LLMs also struggle to utilize parametric knowledge effectively, and processing such knowledge does not ensure accurate elicitation (Allen-Zhu and Li, 2023b).
- **The effect of refusal**: Ideally, LLM should issue a refusal response when external knowledge is irrelevant and lacks corresponding parametric knowledge. However, LLMs tend to

be overconfident in external knowledge regardless of its usefulness (Chen et al., 2024), particularly for cases like $S_3 \sim S_4$ where external knowledge is entirely unhelpful, leading to plausible but incorrect responses (hallucinations). SFT and CT ameliorate this issue. SFT provides examples of refusal responses in the training data, instructing the LLM when to refuse. Meanwhile, CT enhances the LLM's understanding of domain knowledge, improving its ability to judge the efficacy of both external and parametric knowledge in addressing a given question.

- **The effect of knowledge fusion**: When the external knowledge is incomplete, LLMs often struggle to effectively fuse parametric and external information for response generation (Xie et al., 2023). Efficient fusion is generally constrained by factors such as the LLM's knowledge capacity, knowledge boundary perception, noise resistance, and knowledge elicitation ability (Wang et al., 2023c).

Accordingly, to better fuse parametric and external knowledge in LLMs, we identify several key challenges that need addressing. While some work (Allen-Zhu and Li, 2023a,b; Chen et al., 2024; Wang et al., 2023c; Xie et al., 2023) are underway to tackle these issues, the approach from the perspective of knowledge fusion remains underexplored.

- With respect to the noisy information, how to eliminate noise in external knowledge and enhance the noise resistance ability of LLMs, especially in the absence of corresponding parametric knowledge?
- For knowledge infusion, how to optimize the training strategies or methodologies so that the LLM can retain as much knowledge as possible?
- How can LLMs elicit the correct parametric knowledge to answer given questions and accurately recognize its knowledge boundaries, triggering a refusal when neither parametric nor external knowledge is available, rather than generating a hallucinated response?
- How can we optimize the use of parametric and external knowledge to achieve accurate integration when external knowledge is incomplete and the LLM has corresponding default knowledge?

7 Conclusion

This work underscores the nuanced interplay between external and parametric knowledge within LLMs, emphasizing the potential and challenges intrinsic to their fusion. By meticulously deconstructing knowledge fusion into four distinct scenarios and developing a structured pipeline for data construction and knowledge infusion, we have provided a comprehensive examination of LLM behavior across varying contexts of knowledge supplementation. The results indicate that while supervised fine-tuning or enhancing parametric knowledge via continued training is capable of improving the knowledge fusion performance, persistent challenges remain in noise resistance, more effective knowledge infusion, parametric knowledge boundary perception, and accurate knowledge elicitation. These insights lay a foundational framework for future research aimed at achieving a more harmonious and effective synthesis of external and parametric knowledge within LLMs, ultimately advancing their capabilities and applications.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *ArXiv*, abs/2309.14316, 2023a.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. *ArXiv*, abs/2309.14402, 2023b.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *ArXiv*, abs/2305.10403, 2023.

Luca Arrotta, Claudio Bettini, Gabriele Civitarese, and Michele Fiori. Contextgpt: Infusing llms knowledge into neuro-symbolic activity recognition models. *ArXiv*, abs/2403.06586, 2024.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *ArXiv*, abs/2309.16609, 2023.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *ArXiv*, abs/2309.12288, 2023.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*, 2024.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR, 2022.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. *ArXiv*, abs/2309.01431, 2023.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (16):17754–17762, Mar. 2024.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. *ArXiv*, abs/2401.14887, 2024.

Yujuan Ding, Wenqi Fan, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meets llms: Towards retrieval-augmented large language models. *ArXiv*, abs/2405.06211, 2024.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *ArXiv*, abs/2404.17790, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021a.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021b.

Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. Knowledge is a region in weight space for fine-tuned language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1350–1370, 2023.

Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language model inference. *ArXiv*, abs/2301.00303, 2022.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

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

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *ArXiv*, abs/2403.14403, 2024.

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

Fangkai Jiao, Bosheng Ding, Tianze Luo, and Zhanfeng Mo. Panda llm: Training data and evaluation for open-sourced chinese instruction-following large language models. *ArXiv*, abs/2305.03025, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221, 2022.

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

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

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering. *ArXiv*, abs/2203.05115, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. *ArXiv*, abs/2211.05110, 2022.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*, 2024.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. *ArXiv*, abs/2310.01352, 2023.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. We're afraid language models aren't modeling ambiguity. *ArXiv*, abs/2304.14399, 2023a.

Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Recall: A benchmark for llms robustness against external counterfactual knowledge. *ArXiv*, abs/2311.08147, 2023b.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *ArXiv*, abs/2308.08747, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *ArXiv*, abs/2212.10511, 2022.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *ArXiv*, abs/2302.07842, 2023.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. Skill: Structured knowledge infusion for large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588, 2022.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332, 2021.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ArXiv*, abs/2307.06435, 2024.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms. *ArXiv*, abs/2312.05934, 2023.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. Contraqa: Question answering under contradicting contexts. *ArXiv*, abs/2110.07803, 2021.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. Not all contexts are equal: Teaching llms credibility-aware generation. *ArXiv*, abs/2404.06809, 2024.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019.

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. "merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *ArXiv*, abs/2309.08594, 2023.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *ArXiv*, abs/2302.00083, 2023.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *ArXiv*, abs/2307.11019, 2023a.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv*, abs/2307.11019, 2023b.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *ArXiv*, abs/2303.17580, 2023.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ArXiv*, abs/2404.16789, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *ArXiv*, abs/2310.07521, 2023a.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *ArXiv*, abs/2310.00935, 2023b.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315, 2023c.

Yixu Wang, Yan Teng, Kexin Huang, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, and Yingchun Wang. Fake alignment: Are llms really aligned well? *ArXiv*, abs/2311.05915, 2023d.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *ArXiv*, abs/2305.13300, 2023.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *ArXiv*, abs/2401.15884, 2024.

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

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665, 2023a.

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

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *ArXiv*, abs/2310.01558, 2023.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *ArXiv*, abs/2401.01286, 2024.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, 2023a.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, Singapore, 2023b. Association for Computational Linguistics.