Augmented Large Language Models with Parametric Knowledge Guiding

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have significantly advanced natural language processing (NLP) with their impressive language understanding and generation capabilities. However, their performance may be suboptimal for domain-specific tasks that require specialized knowledge due to limited exposure to the related data. Additionally, the lack of transparency of most state-of-the-art (SOTA) LLMs, which can only be accessed via APIs, impedes further fine-tuning with domain custom data. To address these challenges, we propose the novel **Parametric Knowledge Guiding (PKG)** framework, which equips LLMs with a knowledge-guiding module to access relevant knowledge without altering the LLMs' parameters. Our PKG is based on open-source "white-box" language models, allowing offline memory of any knowledge that LLMs require. We demonstrate that our PKG framework can enhance the performance of "black-box" LLMs on a range of domain knowledge-intensive tasks that require factual (+7.9%), tabular (+11.9%), medical (+3.0%), and multimodal (+8.1%) knowledge.

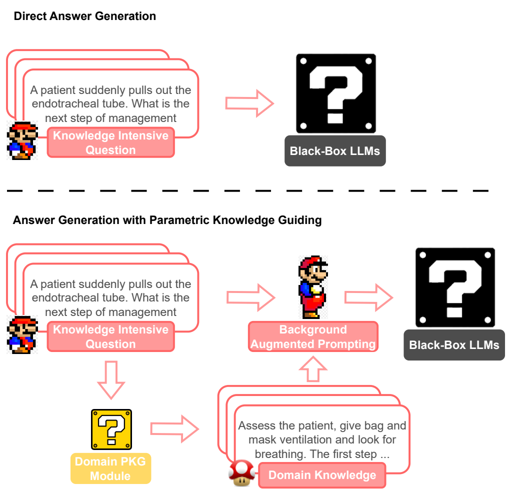


Figure 1: A brief introduction of our parametric knowledge guiding framework (PKG) for augmenting "black box" LLMs on domain knowledge-intensive tasks.

1 Introduction

Large Language Models (LLMs) such as GPT-family (Brown et al., 2020; OpenAI, 2023b) have exhibited impressive proficiency across a diverse range of NLP tasks. These models are typically trained on extensive data from the internet, thereby enabling them to assimilate an immense amount of implicit world knowledge into their parameters. As a result, LLMs have emerged as versatile tools that find numerous applications in both research and industry. For instance, they can be used for machine translation (Jiao et al., 2023), document summarization (Yang et al., 2023), and recommendation systems (Gao et al., 2023). With their exceptional language understanding and generation capabilities, LLMs have opened up new opportunities for diverse industrial applications, such as the recently launched New Bing (Microsoft, 2023) and ChatGPT Plugins (OpenAI, 2023a).

Despite their impressive performance across various general tasks, LLMs may face challenges when applied to domain-specific tasks (Chalkidis, 2023; Kasai et al., 2023; Nascimento et al., 2023) due to their limited exposure to relevant knowledge and vocabulary. Although LLMs acquire implicit world knowledge during pre-training, such knowledge may be insufficient or inappropriate for specific tasks, resulting in less effective performance. Furthermore, many state-of-the-art LLMs are considered "black-box" models, accessible only through APIs. This lack of transparency presents significant challenges and high costs for most researchers and companies seeking to fine-tune these models for their specific use cases or domains. These limitations hinder the adaptability of LLMs to diverse scenarios and domains.

A common approach to enhance LLMs is to leverage retrieval-based methods that access domain-specific knowledge from external sources (Liu, 2022; Shi et al., 2023; Peng et al., 2023a). While these methods have shown promise, they face several challenges. First, they heavily rely on modern dual-stream dense retrieval models (Karpukhin et al., 2020) which suffer from shallow interaction between the query and candidate documents. Second, most dense retrieval models are based on small-scale pre-trained models such as BERT (Devlin et al., 2019) and therefore cannot take advantage of the world knowledge of large-scale pre-trained models. Third, retrieval models may struggle with complex knowledge that requires the integration of information from multiple sources or modalities.

In this work, we propose the **Parametric Knowledge Guiding (PKG)** framework, which enables LLMs to access relevant information without modifying their parameters, by incorporating a trainable background knowledge generation module, as illustrated in Figure 1. Unlike retrieval-based methods, our PKG module utilizes open-source and free-to-use "white-box" language models, LLaMa-7B (Touvron et al., 2023), which encode implicit world knowledge from large-scale pre-training. The framework consists of two steps. First, we train the PKG module with the specific task or domain knowledge via instruction fine-tuning (Ouyang et al., 2022) to capture the necessary expertise. Second, for a given input, the PKG module generates the related knowledge, fed as extra context to the background-augmented prompting for LLMs. By supplying the necessary knowledge, our framework can enhance the performance of LLMs on domain knowledge-intensive tasks.

Our experiments demonstrate that the proposed PKG framework enhances the performance of "black-box" LLMs on various downstream tasks which require domain-specific background knowledge, including factual knowledge (FM2 (Eisenschlos et al., 2021), +7.9%), tabular knowledge (NQ-Table (Herzig et al., 2021), +11.9%), medical knowledge (MedMC-QA (Pal et al., 2022), +3.0%), and multimodal knowledge (ScienceQA (Lu et al., 2022), +8.1%).

We summarize our contributions as follows:

- We propose a novel **Parametric Knowledge Guiding (PKG)** framework that integrates a background knowledge generation module to enhance the performance of LLMs on domain knowledge-intensive tasks.
- We introduce a knowledge-guiding process by first training the parametric modules with specific tasks or domain knowledge and then generating related knowledge as the extra context in the background-augmented prompting.
- We conduct extensive experiments on various downstream tasks to evaluate the effectiveness of our proposed PKG framework. The experiments demonstrate that our PKG framework can improve the capability of LLMs on domain knowledge-intensive tasks.

2 Related Work

Large Language Models. LLMs, such as GPT3 (Brown et al., 2020), Codex (Chen et al., 2021), PaLM (Chowdhery et al., 2022), and GPT4 (OpenAI, 2023b), have gained widespread attention due to their remarkable language understanding and generation capabilities (Wei et al., 2022c; Shi et al., 2022). However, their performance can be limited when it comes to domain-specific tasks, where they may lack exposure to specialized knowledge and vocabulary (Chalkidis, 2023; Kasai et al., 2023; West, 2023). Moreover, while some SOTA LLMs such as InstructGPT3.5 and ChatGPT (Ouyang et al., 2022) exist, they are available only as "black box" APIs due to commercial considerations. This limits researchers and developers with limited resources, who may not be able to access or modify the models' parameters. While open-source LLMs such as OPT-175B (Zhang et al., 2022) and BLOOM-176B (Scao et al., 2022) are available, they lag significantly behind SOTA LLMs on most tasks. Additionally, running and fine-tuning these open LLMs locally requires significant computational resources.

Augmented Large Language Models. ALLMs are a recent popular topic in NLP that aim to enhance the context processing ability of LLMs by incorporating external modules (Mialon et al., 2023; Wu et al., 2023; Shen et al., 2023; Lu et al., 2023; Huang et al., 2023). One approach to achieving this goal is through the use of retrieval-augmented large language models (RLLMs)(Guu et al., 2020; Izacard et al., 2022b; Ram et al., 2023; Shi et al., 2023). RLLMs leverage external knowledge by retrieving relevant documents or passages from knowledge sources using retrieval-based methods such as BM25(Robertson and Zaragoza, 2009) and

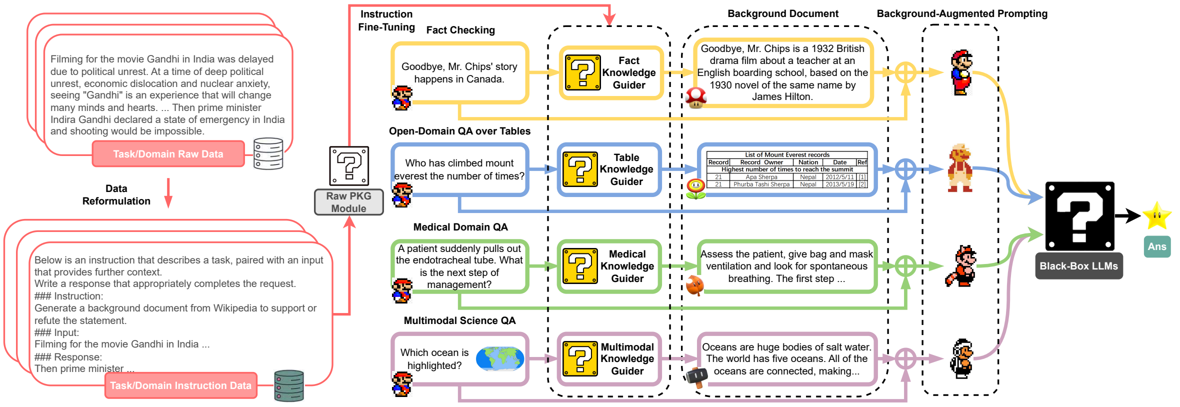


Figure 2: An introduction to the PKG Framework Pipeline: Raw Data Reformulation, PKG Instruction Fine-Tuning and Augmentation of “Black Box" LLMs with Domain-Specific PKG Modules.

DPR (Karpukhin et al., 2020). These retrieved passages are then used as additional contexts to improve the LLMs' performance on the task at hand. Although RLLMs have shown promise in enhancing LLMs' performance, they have certain limitations. For instance, they rely heavily on the dual-stream dense retriever, which leads to shallow interaction between the query and the candidate information. Furthermore, they may struggle with complex queries that require integrating information from multiple sources or modalities.

Instruction Fine-Tuning. IFT is a technique in NLP that aims to align language models with specific user intents (Ouyang et al., 2022). While many LLMs are trained on large datasets of internet data to predict the next word, they may not be tailored to the specific language tasks that users require, meaning that these models are not inherently aligned with their users' needs. Recent research (Wei et al., 2022a; Sanh et al., 2022; Xu et al., 2022; Xie et al., 2022; Xu et al., 2023a; Luo et al., 2023b,a) has highlighted the potential of IFT as a key technique for improving the usability of LLMs. Our proposed approach, PKG, follows the same principle of aligning the basic module with task-specific knowledge to enhance its performance.

3 Parametric Knowledge Guiding for LLMs

In this section, we present our **PKG** framework to guide the reasoning process of LLMs on domain-specific tasks, as shown in Figure 2. These tasks differ from general tasks such as document summarization due to their reliance on specific background knowledge. However, this knowledge may be absent or incomplete in the LLMs' training data. Furthermore, continuous pre-training of LLMs with domain knowledge poses challenges: (1) limited transparency of accessing current SOTA LLMs solely through APIs, and (2) the potentially high fine-tuning cost associated with APIs usage. To tackle these issues, we adhere to the *generate-then-read* paradigm (Yu et al., 2023) and leverage an offline PKG module to generate relevant background knowledge. Our method is first formulated in § 3.1. Next, we describe the background knowledge learning of our PKG modules in § 3.2. Finally, we introduce background-augmented prompting for LLMs in § 3.3.

3.1 Formulation

Given a question/input $\mathcal{Q}$ associated with some contexts, LLMs take the input and generate a response by maximum a posteriori estimation (MAP):

$$\hat{\mathcal{A}} := \text{argmax}_{\mathcal{A}} P(\mathcal{A}|\mathcal{Q}, \mathcal{M}^{LLM}), \quad (1)$$

where $\mathcal{M}^{LLM}$ represents the parameters of the LLMs. However, for tasks that require background knowledge beyond what is contained in the input, such as knowledge-intensive tasks, relying solely on LLMs may not be effective. This is because there may be a significant amount of additional domain-specific knowledge that remains unexploited.

To improve performance, we first introduce an auxiliary PKG module $\mathcal{M}^{PKG}$ to learn specific background knowledge (§3.2). Next, we estimate the input-related background knowledge $\mathcal{K}$ using MAP estimation:

$$\hat{\mathcal{K}} := \text{argmax}_{\mathcal{K}} P(\mathcal{K}|\mathcal{Q}, \mathcal{M}^{PKG}). \quad (2)$$

Finally, the background knowledge $\mathcal{K}$ enriches the input by incorporating background-augmented prompting for LLMs (§ 3.3) in the form:

$$P(\mathcal{A}|\mathcal{Q}) := P(\mathcal{A}|\mathcal{K},\mathcal{Q},\mathcal{M}^{LLM})P(\mathcal{K}|\mathcal{Q},\mathcal{M}^{PKG}). \tag{3}$$

3.2 Background Knowledge Learning

Given a target task or domain, our PKG framework utilizes an open-source language model to learn the relevant knowledge. Figure 2 presents an example of the fact-checking task. This process is divided into two steps. First, we collect raw data about the target task/domain, which serves as our background knowledge. Second, we transform the data into a set of (instruction, input, output) triples. The instruction serves as a prompt for the input and guides the module to learn the expected knowledge.

Next, this set of triples is adopted to tune our basic PKG module with instruction fine-tuning (Ouyang et al., 2022), which optimizes its ability to provide relevant and effective background knowledge to the LLMs. This two-step process can be completed fully offline, without requiring us to provide our data to tune the LLMs. Once trained with the task background knowledge, the PKG module learns to generate domain-specific knowledge to assist the LLMs during runtime.

The instruction data format of the fact-checking task is:

Instruction Format Example

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: {instruction}
Input: {input sentence}
Response: {background}

The `{input sentence}` is a sentence within the specified task. The `{background}` is the background knowledge that the model generates based on the given `{instruction}` and `{input sentence}`. The basic PKG module is trained in a standard supervised way with an auto-regressive manner, where the model generates the `{background}` given the previous context. More instruction data formats for different tasks are presented in Appendix F.

3.3 Background-Augmented Prompting

Instead of directly requesting the LLMs to generate the answer or response for the input question or sentence via APIs, we first instruct the PKG module to generate the background knowledge. In the second step, we utilize the generated background in combination with the input question to derive the final answer from the LLMs. This is similar to the "zero-shot" open-domain question-answering setting that has been widely explored in prior research (Brown et al., 2020; Lazaridou et al., 2022; Yu et al., 2023). The background-augmented prompt of the fact-checking task is:

Background-Augmented Prompt Example

{background}
Claim: {input sentence}
Is the claim true or false?

Finally, the augmented prompt is fed into the LLMs to generate an answer. More prompts for different tasks are presented in Appendix G.

4 Experiment

In this section, we evaluate our proposed PKG framework across four distinct types of knowledge: factual, tabular, medical, and multimodal. Factual knowledge entails the model's ability to access accurate information, serving as a foundational type of knowledge crucial for numerous NLP applications (§ 4.2). Tabular knowledge necessitates the model's capability to access structured knowledge in the form of tables, which is relatively scarce in the training data of LLMs (§ 4.3). Medical knowledge, being highly specialized, exhibits limited exposure within the general data (§ 4.4). Lastly, multimodal knowledge poses a challenge as most LLMs are unable to process non-language information, highlighting the significance of assistance from PKG modules (§ 4.5).

The experimental results depicted in Tables 1 and 2 demonstrate substantial enhancements attained through our PKG framework compared to the baseline systems. These results offer compelling evidence supporting the generalizability and effectiveness of our approach.

4.1 Models Steup

Black-Box LLMs. We adopt one of the SOTA LLM InstructGPT3.5 (Ouyang et al., 2022) as our target "black box" general LLMs, using the

Models	FM2	NQ-Table	MedMC-QA
Direct generation without guiding.			
InstructGPT3.5 (Ouyang et al., 2022)	59.4	16.9	44.4
Generation with retrieval guiding.			
BM25 + InstructGPT3.5 (Karpukhin et al., 2020)	65.2	17.1	-
Contriever + InstructGPT3.5 (Izacard et al., 2022a)	66.0	24.5	-
⋄REPLUG + InstructGPT3.5 (Shi et al., 2023)	65.9	24.3	-
Generation with self-guiding.			
†CoT + InstructGPT3.5 (Kojima et al., 2022)	60.4	21.4	41.5
‡GenRead + InstructGPT3.5 (Yu et al., 2023)	65.5	23.5	44.4
PKG + InstructGPT3.5 (Ours)	**67.3**	**28.8**	**47.4**

Table 1: Evaluating on three different tasks, requiring factual (FM2), tabular (NQ-Table), and medical (MedMC-QA) knowledge. ⋄: we fine-tune the dense retrieval models with the task data. †: we use InstructGPT3.5 to generate the chain-of-thoughts as the background knowledge. ‡: we use InstructGPT3.5 to generate the background documents.

`text-davinic-002` version. With up to 175B parameters, this model is one of the largest LLMs and is pre-trained on a vast amount of internet data, which exhibits great language understanding and generation ability. However, this model can only be accessed through an API, which limits users' interaction.

Basic PKG Module. Our knowledge guiding module employs the open-source and popular foundation model LLaMa-7B (Touvron et al., 2023). It has been pre-trained on massive amounts of text data and possesses extensive world knowledge. Though its performance in many tasks may be inferior to the InstructGPTs, it can be locally fine-tuned and customized (Taori et al., 2023; Xu et al., 2023b; Peng et al., 2023b; Geng et al., 2023), making it an effective starting point for developing a task-specific PKG module.

Baselines. Our work includes three different types of baselines: (1) *Direct generation without guiding*: We do not provide any background knowledge for a given task and ask the InstructGPT to generate the answer or response directly in a zero-shot manner, following the approach of prior works (Brown et al., 2020; Ouyang et al., 2022). (2) *Generation with retrieval guiding*: We follow the retrieve-then-read paradigm (Chen et al., 2017; Yang et al., 2019; Karpukhin et al., 2020) to retrieve related knowledge from external knowledge sources using retrieval models such as BM25 (Robertson and Zaragoza, 2009), DPR (Karpukhin et al., 2020), and Contriever (Izacard et al., 2022a). We fine-tune the DPR on specific tasks following the REPLUG (Shi et al., 2023) method. InstructGPTs then generate responses based on the combination of the question and retrieved background documents. (3) *Generation with self-guiding*: we adopt the InstructGPTs to generate the related background knowledge by themselves with two different methods. The first method, CoT (Kojima et al., 2022), adopts the prompt *"Let's think step-by-step"* to generate the chain-of-thought as the background knowledge. The second method, GenRead (Yu et al., 2023), directly requires the InstructGPTs to provide task-specific knowledge with the prompt *"Please provide the background document from [domain] to [task]."*

4.2 Factual Knowledge

Datasets and Implementation Details. We evaluate our approach on the FM2 dataset (Eisenschlos et al., 2021), which is a benchmark for fact-checking. In this task, given a factual claim, our models are required to determine whether it is true or false. We use the claim in the training set and the corresponding evidence as factual knowledge. Additionally, we sample 100k passages from English Wikipedia, each consisting of up to 256 tokens. We treat the first sentence as the input and the remaining sentences as background knowledge. Accuracy is adopted as the evaluation metric. More details can be found in Appendix A and B.

Results. As shown in Table 1, our PKG outperforms all the baseline systems for fact-checking. In comparison to direct generation, the results reveal

Models	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Base on gpt-3.5-turbo.									
†ChatGPT	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03	78.31
†Chameleon	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53	79.93
Base on text-davinic-002.									
InstructGPT3.5	72.96	62.88	76.09	70.77	62.77	77.84	75.04	65.59	71.66
+CoT	71.94	61.19	74.00	69.50	61.18	75.75	72.61	65.92	70.22
+GenRead	72.91	64.68	76.36	72.14	63.31	76.66	74.96	66.91	72.08
+PKG (Ours)	79.35	82.90	81.91	79.86	74.32	83.41	80.80	80.69	**80.76**

Table 2: Evaluating on the ScienceQA, requiring multimodal science knowledge. †: results from (Lu et al., 2023). `gpt-3.5-turbo` is much more capable than `text-davinic-002`.

that it is necessary to provide extra background knowledge for InstructGPTs with retrieval-based or generation-based methods. Specifically, our PKG outperforms InstructGPT3.5 by 7.9% (67.9% vs. 59.4%), and outperforms REPLUG, a retrieval-based method, by 1.4% (67.3% vs. 65.9%). It is noteworthy that our generation-based method does not necessitate an additional knowledge database as the retrieval-based methods. Additionally, our PKG performs better than the self-guiding method GenRead by 1.8% (67.3% vs. 65.5%), indicating that our PKG can provide more useful information than the InstructGPTs themselves.

4.3 Tabular Knowledge

Datasets and Implementation Details. We evaluate the effectiveness of our approach on the NQ-Table dataset (Herzig et al., 2021), which serves as a benchmark for open-domain question answering over tables. The dataset consists of questions whose answers can be found in a Wikipedia table. We adopted the question in the training set as input and the corresponding flattened table as background knowledge. Our PKG was trained to follow instructions and generate the relevant table. Exact matching is adopted as the evaluation metric. More details can be found in Appendix A and B.

Results. Table 1 demonstrates the superior performance of our PKG framework over all baseline systems on the tabular knowledge-related task. Notably, our PKG outperforms InstructGPT3.5 by a substantial margin of 11.9% (28.8% vs. 16.9%), and outperforms REPLUG, the retrieval-based method, by 4.5% (28.8% vs. 24.3%). Furthermore, our PKG significantly outperforms the self-guiding method GenRead by 5.3% (28.8% vs. 23.5%). These results demonstrate the efficacy and superiority of our approach in leveraging parametric knowledge to augment InstructGPTs for tabular knowledge-related tasks.

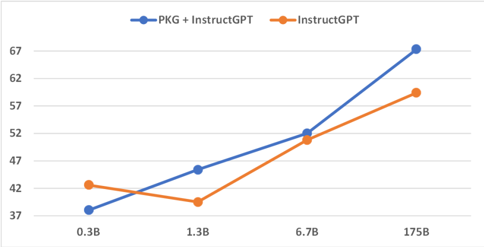
4.4 Medical Knowledge

Datasets and Implementation Details. We evaluate the effectiveness of our approach on the MedMC-QA dataset (Pal et al., 2022), which serves as a benchmark for multi-subject multi-choice medical question answering. Each question requires the use of relevant medical information as background knowledge to provide the correct answer. We use the questions in the training set as input and the corresponding medical explanation as background knowledge. Our PKG is trained to follow the instruction and generate the relevant medical background. Accuracy is the evaluation metric. Unlike the previous tasks with all Wikipedia passages as the knowledge database, we do not have access to an external medical knowledge database, and thus we do not evaluate the performance of retrieval-based methods on this task. More details can be found in Appendix A and B.

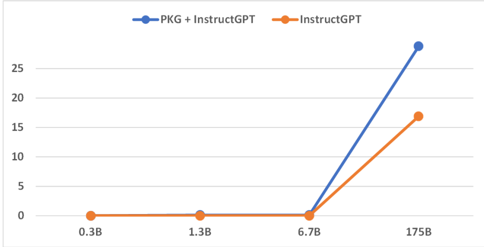
Results. Our PKG framework also outperforms all baseline systems on this medical knowledge-related task, as shown in Table 1. Specifically, our PKG outperforms InstructGPT3.5 by 3.0% (47.4% vs. 44.4%). It is worth noting that the baseline self-guiding methods, CoT and GenRead, do not improve the performance of InstructGPTs. This may be due to the fact that InstructGPTs lack sufficient medical information to effectively solve this task.

4.5 Multimodal Knowledge

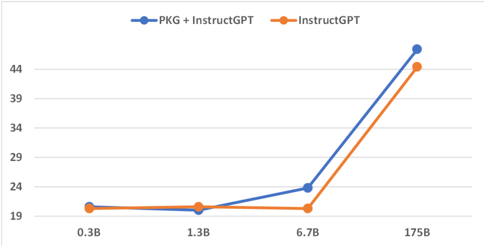
Datasets and Implementation Details. Our approach is evaluated on the ScienceQA dataset (Lu



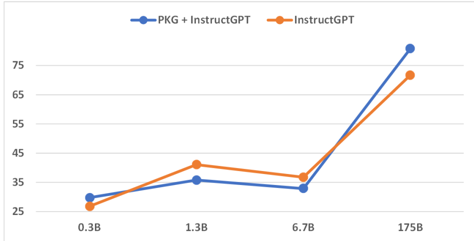
(a) Accuracy on FM2.



(b) Exact Matching on NQ-Table.



(c) Accuracy on MedMC-QA.



(d) Average Accuracy on ScienceQA

Figure 3: Comparing our PKGs framework with the direct generation on various types of InstructGPT. The number indicates the number of parameters in the InstructGPT. 0.3B: text-ada-001, 1.3B: text-babbage-001, 6.7B: text-curie-001, 175B: text-davinci-002.

et al., 2022), which presents a challenging multimodal multiple-choice question-answering task covering diverse science topics. Each question requires leveraging relevant scientific background knowledge to provide the correct answer. We use the training set's questions as input and their corresponding science lecture as background knowledge. To handle the images information, we augment our basic PKG module with the CLIP-ViT (Radford et al., 2021) to extract visual features, which are then fused with text features using a simple one-head cross-attention mechanism in each layer of LLaMa:

$$\mathcal{H} := \mathcal{H}^{txt} + \mathcal{W}^{o} \left(\text{softmax}\left((\mathcal{W}^{q}\mathcal{H}^{txt})(\mathcal{W}^{k}\mathcal{H}^{img})^{T}\right)(\mathcal{W}^{v}\mathcal{H}^{img})\right), \tag{4}$$

where $\mathcal{W}^{o,q,k,v}$ are the linear projection, $\mathcal{H}^{txt,img}$ are the hidden states of texts and images. We adopt accuracy as the evaluation metric. More details can be found in Appendix A and B.

Similarly, this task is also difficult to obtain an external multimodal science knowledge database, retrieval-based methods are not considered. To facilitate a fair comparison of our methods, we include two additional baseline systems (Lu et al., 2023) based on the `gpt-3.5-turbo` model. The first baseline is ChatGPT direct generation, and the second is the Chameleon model, which utilizes several external tools, such as searching, OCR, and captioning. According to OpenAI, the `gpt-3.5-turbo` model is more capable than `text-davinic-002` (OpenAI, 2023c).

Size	FM2	NQ-Table	MedMC-QA	SciQA
7B	**67.3**	**28.8**	**47.4**	**80.8**
2.7B	59.6	17.9	34.4	79.5
1.3B	58.2	16.5	33.9	77.0
0.3B	56.4	14.6	31.7	68.7

Table 3: Comparing various sizes of language models as the basic PKG modules.

Results. Table 2 shows that our PKG framework achieves a significant improvement in the performance of InstructGPTs on the multimodal scientific knowledge-related task. Specifically, the average accuracy is increased by 9.1% (80.76% vs. 71.66%), demonstrating the effectiveness of our approach. In contrast, other guiding methods, CoT (-1.44%) and GenRead (+0.42%), hard to improve the performance of InstructGPTs. Moreover, our PKG framework outperforms the `gpt-3.5-turbo` based models on average by 2.45% (80.76% vs. 78.31%), despite using weaker InstructGPTs.

4.6 Analysis

Scale of LLMs. Figure 3 presents the impact of our PKG framework on several "black-box" LMs, including `text-ada-001`, `text-babbage-001`, `text-curie-001`, and `text-davinci-002`. The

<table>
<tr><td>Input (FM2)</td><td>The Natural is a book about Roy Hobbs a natural southpaw boxer who goes on to win the heavyweight title from Boom Boom Mancini.</td></tr>
<tr><td>PKG Guiding</td><td>This is a story about a baseball player, Roy Hobbs, who is shot by a woman, Harriet Bird, and then recovers to become a great player.</td></tr>
<tr><td>Input (NQ-Table)</td><td>Batman The Enemy Within episode 5 release date</td></tr>
<tr><td>PKG Guiding</td><td>Batman: The Enemy Within
<table>
<tr><th>Ep</th><th>Title</th><th>Directed by</th><th>Release date</th></tr>
<tr><td>1</td><td>"The Enigma"</td><td>Kent Mudle</td><td>August 8, 2017</td></tr>
<tr><td>2</td><td>"The Pact"</td><td>Matthew Leach</td><td>October 3, 2017</td></tr>
<tr><td>3</td><td>"Fractured Mask"</td><td>Sean Manning</td><td>November 21, 2017</td></tr>
<tr><td>4</td><td>"What Ails You"</td><td>Chris Rieser</td><td>January 23, 2018</td></tr>
<tr><td>5</td><td>"Same Stitch"</td><td>Kent Mudle</td><td>March 27, 2018</td></tr>
</table></td></tr>
<tr><td>Input (MedMC-QA)</td><td>Calcium ions triggers muscle contraction by binding to: Options: (A) Actin (B) Myosin (C) Troponin (D) Tropomyosin</td></tr>
<tr><td>PKG Guiding</td><td>Troponin is a complex of three proteins that are attached to actin. Calcium ions bind to troponin-C, which then causes the troponin-I and troponin-T to shift and allow the tropomyosin to move off of the active site of the myosin so that the myosin head can then bind to actin and cause contraction.</td></tr>
</table>

Table 4: Examples of background documents generated by our PKGs to guide different tasks. Clues to answering the input are highlighted in blue within the documents.

results suggest that the effectiveness of our approach is correlated with the size of the LMs, with larger LMs benefiting more from our PKGs than smaller ones. Specifically, in Figure 3b, the small LMs show negligible exact matching scores on the tabular task, with or without the background knowledge from our PKGs, while the LLMs exhibit significantly better performance. In Figure 3c, the 0.3B and 1.3B LMs perform similarly on the medical domain task, while the 6.7B LM shows improved performance with the additional knowledge. This difference can be attributed to the relatively weaker language understanding capabilities of smaller LMs, which struggle to reason over contexts and generate the correct responses even with relevant knowledge from our PKGs. These observations align with the emergent abilities of LLMs, as discussed in (Wei et al., 2022b). Therefore, the scale of LLMs is a critical factor for achieving better performance.

Scale of PKGs. We conducted an investigation of various sizes of language models as basic PKG modules in Table 3. Since LLaMa-7B is the smallest model in the LLaMa family, we conducted experiments on the OPT family (Zhang et al., 2022), another open-source large-scale language model with a similar structure to LLaMa. Our observations reveal that larger basic PKGs tend to exhibit superior performance. For example, increasing the number of parameters from 1.3B to 2.7B leads to performance improvements of 1.4% on FM2, 1.4% on NQ-Table, 0.5% on MedMC-QA, and 2.5% on ScienceQA, which is consistent with the scaling law (Kaplan et al., 2020).

Examples of Generated Background Documents. Table 4 presents examples of background documents generated by our PKGs to assist LLMs in different tasks. For the factual task, our PKG can supply input-related factual information to support or refute the input, such as the example of Roy Hobbs being a baseball player and not a boxer. For the tabular task, our PKG can offer an input-related background table, like the episode table of Batman. For the medical task, our PKG can provide relevant medical knowledge, such as the background of calcium ions. Since the space is not enough, examples for the multimodal tasks and additional examples can be found in Appendix D.

5 Conclusion

In this work, we propose the novel **Parametric Knowledge Guiding (PKG)** framework to enhance the performance of “black-box" LLMs on domain-specific tasks by equipping them with a knowledge-guiding module. Our approach allows for access to relevant knowledge at runtime without altering the “black-box" LLM’s parameters. The extensive experiments demonstrate the effectiveness of our PKG framework for various domain knowledge-intensive tasks.

Limitations

Although our PKGs have shown strong performance on the presented datasets, they may still suffer from hallucination errors, leading to the provision of incorrect background knowledge. We provide examples of such errors in Appendix E. Combining our approach with retrieval methods to enhance generative faithfulness is a promising direction for future research.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Ilias Chalkidis. 2023. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan L. Boyd-Graber. 2021. Fool me twice: Entailment from wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 352–365. Association for Computational Linguistics.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *CoRR*, abs/2303.14524.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 512–519. Association for Computational Linguistics.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. *CoRR*, abs/2304.12995.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

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

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating GPT-4 and chatgpt on japanese medical licensing examinations. *CoRR*, abs/2303.18027.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

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

Jerry Liu. 2022. LlamaIndex.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *CoRR*, abs/2209.09513.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *CoRR*, abs/2304.09842.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *CoRR*, abs/2308.09583.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023b. Wizardcoder: Empowering code large language models with evol-instruct. *CoRR*, abs/2306.08568.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *CoRR*, abs/2302.07842.

Microsoft. 2023. New bing. Webpage. Accessed on May 8, 2023.

Castro Nascimento, Cayque Monteiro, Pimentel, and André Silva. 2023. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655. PMID: 36926868.

OpenAI. 2023a. Chatgpt plugins. Webpage. Accessed on May 8, 2023.

OpenAI. 2023b. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2023c. Models overview. Webpage. Accessed on May 8, 2023.

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023b. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *CoRR*, abs/2210.03057.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. *CoRR*, abs/2301.12652.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *CoRR*, abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Colin G. West. 2023. AI and the FCI: can chatgpt project an understanding of introductory physics? *CoRR*, abs/2303.01067.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 602–631. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023b. Wizardlm: Empowering large language models to follow complex instructions.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Zeroprompt: Scaling prompt-based pretraining to 1, 000 tasks improves zero-shot generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4235–4252. Association for Computational Linguistics.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 72–77. Association for Computational Linguistics.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *CoRR*, abs/2302.08081.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

A Datasets and Splits

Our experiments include four different benchmarks to evaluate our PKG framework:

- **Fool Me Twice (FM2)** (Eisenschlos et al., 2021) is a fact-checking task, which contains a set of claims with evidence that were originally scarped from Wikipedia.
- **Natural Questions Over Tables (NQ-Table)** (Herzig et al., 2021) is an open-domain question-answering task over table knowledge, which is mined from real Google search queries, and the answers are spans in Wikipedia tables identified by human annotators.
- **Multi-Subject Multi-Choice Dataset for Medical domain (MedMC-QA)** (Pal et al., 2022) is a medical question-answering task, which contains a set of real-world medical entrance exam questions and answers.
- **Multimodal Reasoning for Science Question Answering (ScienceQA)** (Lu et al., 2022) is a multimodal reasoning task, which consists of multimodal multiple-choice questions with a diverse set of science topics.

In Table 5, we show the dataset splits and statistics.

B Implementation Details

We employ LLaMa-7B (Touvron et al., 2023) as the backbone models for implementing the PKG modules. The AdamW optimizer is used, with 10% warmup steps. Training of the PKG modules is performed on 8 V100 GPUs. The vision encoder for ScienceQA is CLIP-ViT-B/32, whose parameters are not updated during training. In our experiments, we extensively utilize the open-source code *LLaMa-X*.[1] For more specific implementation details, please refer to Table 6.

We implement other baseline methods based on the following repositories:

- BM25 + GPT3.5: `https://github.com/castorini/pyserini`
- REPLUG + GPT3.5: `https://github.com/facebookresearch/DPR/tree/main`
- CoT + GPT3.5: `https://github.com/kojima-takeshi188/zero_shot_cot`

[1] `https://github.com/AetherCortex/Llama-X`

Datasets	Domain	Train	Valid	Test	Test labels
FM2 (Eisenschlos et al., 2021)	Factual	10,419	1,169	1,380	Public
NQ-Table (Herzig et al., 2021)	Tabular	9,594	1,068	959	Public
MedMC-QA (Pal et al., 2022)	Medical	160,869	4,183	6,150	Private
ScienceQA (Lu et al., 2022)	Multimodal	12,726	4,241	4,241	Public

Table 5: Datasets splits and statistics. For MedMC-QA, labels in the test are hidden, so the model performance is evaluated on the validation set.

Settings	FM2	NQ-Table	MedMC-QA	ScienceQA
Peak learning rate	2e-5	2e-5	2e-5	2e-5
β_1, β_2	[0.9,0.999]	[0.9,0.999]	[0.9,0.999]	[0.9,0.999]
ϵ	1e-8	1e-8	1e-8	1e-8
Weight decay	0	0	0	0
Total batch size	64	32	32	32
Total training epochs	3	10	3	5
Warmup Schedule	cosine	cosine	cosine	cosine
Warmup ratio	0.1	0.1	0.1	0.1
Precision	fp16	fp16	fp16	fp16

Table 6: Hyperparameters settings of our PKG modules on different tasks.

- GenRead + GPT3.5: https://github.com/wyu97/GenRead

C All Results of Figure 3 in the Main Paper

In Figure 3 of the main paper, we compare our PKGs framework with the direct generation on various types of LMs. We include all results in Table 7.

D Case Studies

Additional examples of background documents generated by our baseline methods (CoT and GenRead) and PKGs for different tasks are presented in Table 8, Table 9, Table 10, and Table 11. These examples highlight how our PKGs can provide valuable information to assist LLMs in answering specific questions. Furthermore, Table 12 compares our PKGs with retrieval-based methods, demonstrating that the retrieval methods are unable to offer relevant background documents to address the given question effectively.

E Errors

Table 13 showcases examples of hallucination errors generated by our PKGs. Similar to other LLMs, our PKGs may introduce fabricated background knowledge in certain instances.

F Instruction Formats

- FM2:

```
  Below is an instruction that describes a task, paired with
an input that provides further context.
Write a response that appropriately completes the request.
### Instruction:
Generate a background document from Wikipedia to support or
refute the statement.
### Input:
Statement: xxx
### Response:
<background fact>
```

- NQ-Table:

```
  Below is an instruction that describes a task, paired with
an input that provides further context.
Write a response that appropriately completes the request.
### Instruction:
Generate a background table from Wikipedia to answer the given
question.
### Input:
Question: xxx
### Response:
<background table>
```

- MedMC-QA

```
  Below is an instruction that describes a task, paired with
an input that provides further context.
```

Methods	FM2	NQ-Table	MedMC-QA	ScienceQA
PKG-Davinci	**67.3**	**28.8**	**47.4**	**80.76**
PKG-Curie	52.0	0.1	23.8	32.87
PKG-Babbage	45.4	0.1	20.0	35.77
PKG-Ada	38.0	0.0	20.6	29.76
Direct-Davinci	59.4	16.9	44.4	71.66
Direct-Curie	50.8	0.0	20.3	36.76
Direct-Babbage	39.5	0.0	20.6	41.08
Direct-Ada	42.6	0.0	20.3	26.79

Table 7: All experiments results of Figure 3 for different sizes of LMs in the main paper.

```
Write a response that appropriately completes the request.
### Instruction:
Generate a background document from the medical domain to answer the given question.
### Input:
Question: xxx
### Response:
<background medical knowledge>
```

- ScienceQA: We follow the "QCM-LE" format in MM-CoT (**?**), where "Q" is the question, "C" is the context, "M" is the choices, "L" is the lecture and "E" is the explanation. Please refer to the paper of MM-CoT for more details.

G Background-Augmented Prompts

- FM2:

```
<background>

claim: <query>

Is the claim true or false?
```

- NQ-Table:

```
Refer to the background below and answer the following question with just a few words. The answer should be less than 5 words.

Background: <background>

Question: <question>

Answer:
```

- MedMC-QA:

```
Refer to the medical background below and answer the following question.

Background: <background>

Question: <question>
Options: <options>

Please only choose the answer from options. The answer is:
```

- ScienceQA:

```
Question: <question>
BECAUSE: <background>
Options: <options>
Please only choose the answer from options. The answer is:
```

Input (FM2)	The Natural is a book about Roy Hobbs a natural southpaw boxer who goes on to win the heavyweight title from Boom Boom Mancini.
CoT Guiding	The first part of the claim is that The Natural is a book. This is true - The Natural is a novel written by Bernard Malamud. The second part of the claim is that the book is about Roy Hobbs, a natural southpaw boxer. This is also true - the novel follows the character of Roy Hobbs, a talented young baseball player who is drafted by a major league team. The third part of the claim is that Hobbs goes on to win the heavyweight title from Boom Boom Mancini. This is false - in the novel, Hobbs does not win the heavyweight title from Boom Boom Mancini.
GenRead Guiding	The Natural is a 1984 American sports drama film based on Bernard Malamud's 1952 baseball novel of the same name, directed by Barry Levinson and starring Robert Redford, Glenn Close, and Robert Duvall. The film, like the book, recounts the experiences of Roy Hobbs, an extraordinarily talented left-handed pitcher, who becomes a late-blooming professional baseball player with the New York Knights.
PKG Guiding	This is a story about a baseball player, Roy Hobbs, who is shot by a woman, Harriet Bird, and then recovers to become a great player.
Input (FM2)	Goodbye, Mr. Chips' story happens in Canada.
CoT Guiding	First, let's recall what the story is about. The story is about a schoolteacher, Mr. Chips, who spends his entire life teaching at a boys' boarding school in England. Now, let's think about whether the story could have taken place in Canada. Could a schoolteacher in Canada spend his entire life teaching at a boys' boarding school? It's possible, but it seems unlikely. Could a schoolteacher in Canada spend his entire life teaching at any kind of school? Yes, that's possible. Therefore, the claim that the story happens in Canada is false.
GenRead Guiding	Goodbye, Mr. Chips is a novel by James Hilton, first published in 1934. The novel tells the story of a schoolteacher, Mr. Chipping, who is nicknamed "Chips". The novel was Hilton's best-selling book of all time, and was adapted into two films and a television series. The novel is set in the fictional public school of Brookfield, in the English county of Leicestershire. However, in the film adaptations, the school is moved to the Canadian Rockies.
PKG Guiding	Goodbye, Mr. Chips is a 1934 British drama film based on the 1934 novel of the same name by James Hilton, which tells the story of a school headmaster at an English boys' school from 1870 to 1935.

Table 8: Examples of background documents generated by our baseline methods and PKGs for FM2. Clues to answering the input are highlighted in blue within the documents. The information that will disrupt LLMs is highlighted in red.

<table>
<tr><td>Input (NQ-Table)</td><td>Batman The Enemy Within episode 5 release date</td></tr>
<tr><td>CoT Guiding</td><td>1. When is episode 5 of Batman the enemy released? 2. What is the release date of episode 5 of Batman the enemy? The release date of episode 5 of Batman the enemy is October 10, 2016.</td></tr>
<tr><td>GenRead Guiding</td><td>"Batman: The Enemy Within" is an episodic point-and-click graphic adventure video game developed and published by Telltale Games and distributed by Warner Bros. Interactive Entertainment under its DC Entertainment label. The game is a sequel to 2016's "Batman: The Telltale Series", and was released for Microsoft Windows, macOS, PlayStation 4, Xbox One, Android, and iOS. The first episode was released on August 8, 2017, with the fifth and final episode released on March 27, 2018.</td></tr>
<tr><td>PKG Guiding</td><td>Batman: The Enemy Within
<table>
<tr><th>Ep</th><th>Title</th><th>Directed by</th><th>Release date</th></tr>
<tr><td>1</td><td>"The Enigma"</td><td>Kent Mudle</td><td>August 8, 2017</td></tr>
<tr><td>2</td><td>"The Pact"</td><td>Matthew Leach</td><td>October 3, 2017</td></tr>
<tr><td>3</td><td>"Fractured Mask"</td><td>Sean Manning</td><td>November 21, 2017</td></tr>
<tr><td>4</td><td>"What Ails You"</td><td>Chris Rieser</td><td>January 23, 2018</td></tr>
<tr><td>5</td><td>"Same Stitch"</td><td>Kent Mudle</td><td>March 27, 2018</td></tr>
</table></td></tr>
<tr><td>Input (NQ-Table)</td><td>the old man and the sea page count</td></tr>
<tr><td>CoT Guiding</td><td>The old man is fishing and he catches a big fish. He struggles to bring the fish in, and finally does. Now, he has to get the fish home. So, he ties it to his boat and starts rowing. The old man is rowing for a long time and finally makes it back to shore. He drags the fish up onto the beach and starts to cut it open. Now, we're at the part where the old man is cutting open the fish. He's cut open the belly and is starting to pull out the guts. He's got a knife in one hand and is pulling out the guts with the other. The old man is still cutting open the fish. He's almost done now. He's got the knife in one hand and the guts in the other. He's about to pull out the last of the guts when he suddenly feels something tugging on his line. The old man has been pulled out to sea by the big fish! He's struggling to get back to shore, but the fish is too strong. The old man is fighting for his life now. The old man is still struggling to get back to shore. He's getting tired now, and the fish is still pulling him out to sea. The old man knows he's going to die soon. The old man finally dies and is pulled out to sea by the big fish.</td></tr>
<tr><td>GenRead Guiding</td><td>The Old Man and the Sea is a novel by Ernest Hemingway that was first published in 1952. The novel tells the story of an aging Cuban fisherman who struggles to catch a giant marlin. The Old Man and the Sea was published in 1952 in Life magazine. The novel was adapted into a 1958 film of the same name, starring Spencer Tracy. The Old Man and the Sea won the Pulitzer Prize for Fiction in 1953.</td></tr>
<tr><td>PKG Guiding</td><td>The Old Man and the Sea
<table>
<tr><th>Author</th><th>Language</th><th>Genre</th><th>Pages</th></tr>
<tr><td>Ernest Hemingway</td><td>English</td><td>Literary Fiction</td><td>127</td></tr>
</table></td></tr>
</table>

Table 9: Examples of background documents generated by our baseline methods and PKGs for NQ-Table. Clues to answering the input are highlighted in blue within the documents. The information that will disrupt LLMs is highlighted in red.

Input (MedMC-QA)	Calcium ions triggers muscle contraction by binding to: Options: (A) Actin (B) Myosin (C) Troponin (D) Tropomyosin (Correct Answer: C)
CoT Guiding	First, we need to understand what a muscle contraction is. A muscle contraction is when the muscle fibers in the muscle shorten. This is caused by the sliding of the actin and myosin filaments past each other. Next, we need to understand what role calcium ions play in muscle contraction. Calcium ions are needed for the muscle contraction to occur because they are needed to trigger the release of the myosin head from the actin filament. So, the answer is (C) Troponin.
GenRead Guiding	Calcium ions play an important role in muscle contraction. When calcium ions bind to troponin, they cause a conformational change in the protein that allows myosin and actin to interact. This interaction between myosin and actin is what causes muscle contraction.
PKG Guiding	Troponin is a complex of three proteins that are attached to actin. Calcium ions bind to troponin-C, which then causes the troponin-I and troponin-T to shift and allow the tropomyosin to move off of the active site of the myosin so that the myosin head can then bind to actin and cause contraction.
Input (MedMC-QA)	Which of the following statements is true regarding medical aboion? Options: (A) Ultrasound should be done in all cases (B) If the patient has an IUCD in-situ, it doesn't need to be removed (C) Can only be done up to 72 days (D) Only a person ceified under MTP act can perform medical termination of pregnancy (Correct Answer: D)
CoT Guiding	First, we need to understand what medical abortion is. Medical abortion is the use of medication to end a pregnancy. Next, we need to consider the options given. Option A is not necessarily true - ultrasound may not be needed in all cases, but it can be helpful in determining how far along the pregnancy is. Option B is also not necessarily true - in some cases, the IUCD (intrauterine contraceptive device) may need to be removed before the medical abortion can be performed. Option C is generally true - medical abortions can usually be performed up to 72 days after the last menstrual period. Option D is also generally true - only a certified healthcare provider can perform a medical abortion.
GenRead Guiding	Medical abortion is a safe and effective way to end a pregnancy. It can be done up to 72 days after the first day of the last menstrual period. A medical abortion can be done with pills or with a shot.
PKG Guiding	Medical aboion can be done up to 9 weeks and 6 days. Ultrasound is not required for medical aboion. If the patient has an IUCD in-situ, it should be removed. Only a person ceified under MTP act can perform medical termination of pregnancy.

Table 10: Examples of background documents generated by our baseline methods and PKGs for MedMC-QA. Clues to answering the input are highlighted in blue within the documents. The information that will disrupt LLMs is highlighted in red.

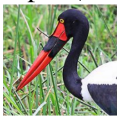
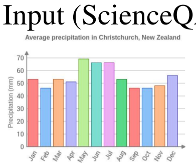
Input (ScienceQA) 	Which animal's neck is also adapted for hunting prey while keeping the rest of its body still? Context: Saddle-billed storks live near wetlands and lakes. They eat mostly fish. The 's neck helps it grab fish while keeping the rest of its body still. If the stork had to move its body, it might scare the fish away. Figure: saddle-billed stork. A black and white bird is standing in the grass. Options: (A) northern pintail (B) black-headed heron (Correct Answer: B)
CoT Guiding	Option A is incorrect because northern pintails are ducks and do not have necks adapted for hunting prey. Option B is correct because black-headed herons are wading birds that use their long necks to help them catch fish.
GenRead Guiding	The answer to the question is option (B), the painted stork. The painted stork's neck is adapted for hunting prey while keeping the rest of its body still, which helps it grab fish while minimizing movement that could scare the fish away.
PKG Guiding	Look at the picture of the saddle-billed stork. The saddle-billed stork has a long neck. Its neck is adapted for hunting prey while keeping the rest of its body still. Now look at each animal. Figure out which animal has a similar adaptation. The black-headed heron has a long neck. Its neck is adapted for hunting prey while keeping the rest of its body still. The northern pintail has a short neck. Its neck is not adapted for hunting prey while keeping the rest of its body still.
Input (ScienceQA) 	Which month is the wettest on average in Christchurch? Context: Use the graph to answer the question below. A series of photographs showing a variety of colorful objects. Options: (A) April (B) December (C) May (Correct Answer: C)
CoT Guiding	The wettest month on average in Christchurch is December.
GenRead Guiding	The graph shows that December is the wettest month on average in Christchurch.
PKG Guiding	Scientists record data from places around the world. Precipitation, or rain and snow, is one type of climate data. A bar graph can be used to show the average amount of precipitation each month. Months with taller bars have more precipitation on average. To describe the average precipitation trends in Christchurch, look at the graph. Choice "Apr" is incorrect. Choice "May" is incorrect. Choice "Dec" is incorrect. May has an average monthly precipitation of about 70 millimeters. This is higher than in any other month. So, May is the wettest month on average.

Table 11: Examples of background documents generated by our baseline methods and PKGs for ScienceQA. Clues to answering the input are highlighted in blue within the documents. The information that will disrupt LLMs is highlighted in red.

Input (NQ-Table)	Batman The Enemy Within episode 5 release date
BM25 Retrieved	is either visited by Bruce or decides to become the hero's archenemy. However, if he was a criminal, he is shown playing with a doll version of Bruce, which he promises to see again. All episodes below were released for Windows, macOS, PlayStation, Xbox One, and mobile platforms on the dates given. The Nintendo Switch version was released as a single package on October 2, 2018. "Batman: The Enemy Within" was considered to be an improvement over its predecessor, earning praise for its story, choices, action sequences, and portrayal of the Batman mythos. However, the presence of technical issues, and
REPLUG Retrieved	Babylon 5: The Legend of the Rangers Babylon 5: The Legend of the Rangers (subtitled: To Live and Die in Starlight) is the fifth telefilm set in the "Babylon 5" universe (not including the pilot,). Originally airing January 19, 2002 on the Sci Fi Channel, it was written by J. Michael Straczynski and directed by Mike Vejar. Though shot as a pilot for a possible new series, it aired opposite NFL playoffs and the subsequent poor ratings led to it not being picked up. As the Shadow War ended, hundreds of civilizations were devastated. It is up to the
PKG Guiding	Batman: The Enemy Within Ep | Title | Directed by | Release date 1 | "The Enigma" | Kent Mudle | August 8, 2017 2 | "The Pact" | Matthew Leach | October 3, 2017 3 | "Fractured Mask" | Sean Manning | November 21, 2017 4 | "What Ails You" | Chris Rieser | January 23, 2018 5 | "Same Stitch" | Kent Mudle | March 27, 2018
Input (NQ-Table)	the old man and the sea page count
BM25 Retrieved	by Magneto's Sentinels for plotting against Magneto. In "JLA/Avengers", Count Nefaria is seen in #4 among the other villains enthralled by Krona to defend his stronghold. He is shown fighting Superman in a panel spreading across two-pages. In the pages of "Old Man Logan", the elderly Logan awoke on Earth-616 and had a flashback to where Count Nefaria, Red Skull, Baron Blood, Spiral, and Whirlwind were standing over the dead bodies of the superheroes the day when the villains rose and the heroes fell. Count Nefaria appears in the "What If?" story "What If the X-Men Died on their First
REPLUG Retrieved	The Old Man and the Sea The Old Man and the Sea is a short novel written by the American author Ernest Hemingway in 1951 in Cuba, and published in 1952. It was the last major work of fiction by Hemingway that was published during his lifetime. One of his most famous works, it tells the story of Santiago, an aging Cuban fisherman who struggles with a giant marlin far out in the Gulf Stream off the coast of Cuba. In 1953, "The Old Man and the Sea" was awarded the Pulitzer Prize for Fiction, and it was cited by
PKG Guiding	The Old Man and the Sea Author | Language | Genre | Pages Ernest Hemingway | English | Literary Fiction | 127

Table 12: Examples of background documents generated by our baseline methods and PKGs for NQ-Table. Clues to answering the input are highlighted in blue within the documents.

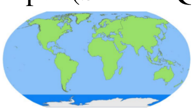
<table>
<tr><td>Input (FM2)</td><td>Hadrian started building a wall that he was never able to complete. (Correct Answer: True)</td></tr>
<tr><td>PKG Guiding</td><td>Either Hadrian or his successor Antoninus Pius started the wall’s construction.</td></tr>
<tr><td>Input (NQ-Table)</td><td>who won game 4 of the 2000 nba finals (Correct Answer: Lakers)</td></tr>
<tr><td>PKG Guiding</td><td>2000 NBA Finals
<table>
<tr><th>Game</th><th>Home Team</th><th>Result</th><th>Road Team</th></tr>
<tr><td>Game 1</td><td>Los Angeles Lakers</td><td>116-86</td><td>Portland Trail Blazers</td></tr>
<tr><td>Game 2</td><td>Los Angeles Lakers</td><td>100-86</td><td>Portland Trail Blazers</td></tr>
<tr><td>Game 3</td><td>Portland Trail Blazers</td><td>86-80</td><td>Los Angeles Lakers</td></tr>
<tr><td>Game 4</td><td>Portland Trail Blazers</td><td>89-78</td><td>Los Angeles Lakers</td></tr>
</table></td></tr>
<tr><td>Input (MedMC-QA)</td><td>Reciprocal arm taper in Options: (A) 1 dimension (B) 2 dimension (C) 3 dimension (D) Not tapered (Correct Answer: A)</td></tr>
<tr><td>PKG Guiding</td><td>Reciprocal arm taper is seen in 3 dimension.</td></tr>
<tr><td>Input (ScienceQA)
</td><td>Which ocean is highlighted? Context: A painting of a penguin on a blue background. Options: (A) the Atlantic Ocean (B) the Indian Ocean (C) the Southern Ocean (D) the Arctic Ocean (Correct Answer: C)</td></tr>
<tr><td>PKG Guiding</td><td>Oceans are huge bodies of salt water. The world has five oceans. All of the oceans are connected, making one world ocean. This is the Pacific Ocean.</td></tr>
</table>

Table 13: Examples of hallucination errors. red: indicates the errors.