Retrieval-Augmented Visual Question Answering via Built-in Autoregressive Search Engines

Xinwei Long[1], Zhiyuan Ma[1], Ermo Hua[1], Kaiyan Zhang[1], Biqing Qi[2], Bowen Zhou[1]*

[1]Department of Electronic Engineering, Tsinghua University
[2]Shanghai Artificial Intelligence Laboratory
longxw22@mails.tsinghua.edu.cn

Abstract

Retrieval-augmented generation (RAG) has emerged to address the knowledge-intensive visual question answering (VQA) task. Current methods mainly employ separate retrieval and generation modules to acquire external knowledge and generate answers, respectively. We propose ReAuSE, an alternative to the previous RAG model for the knowledge-based VQA task, which seamlessly integrates knowledge retriever into the generative multi-modal large language model, serving as a built-in search engine. Specifically, our model functions both as a generative retriever and an accurate answer generator. It not only helps retrieve documents from the knowledge base by producing identifiers for each document, but it also answers visual questions based on the retrieved documents. Furthermore, we also propose a reinforced retrieval calibration module from relevance feedback to improve retrieval performance and align with the preferences for accurate answer generation. Extensive experiments on two representative OKVQA and A-OKVQA datasets demonstrate significant improvements ranging from 2.9% to 9.6% across all evaluation metrics when compared to strong baselines.

Introduction

The Visual Question Answering (VQA) task aims to answer questions based on a user-provided image, which has received significant attention from CV and NLP community (Antol et al. 2015; Hu et al. 2017; Shen et al. 2023b; Sun et al. 2024; Zhu et al. 2024). Early VQA methods (Mascharka et al. 2018; Gao et al. 2019) mainly focus on understanding visual elements within the image. Recently, the research trend of VQA has shifted towards knowledge-intensive scenarios (Shah et al. 2019), requiring the incorporation of external knowledge and joint reasoning over multi-modal content to generate accurate answers. However, existing methods generally face challenges in effectively acquiring relevant information from large-scale knowledge bases using multi-modal queries (Lin et al. 2023).

Retrieval-augmented generation (RAG) (Chan et al. 2024; Chen et al. 2024) has recently emerged as a promising approach for knowledge-based visual question answering (KBVQA) tasks (Gao et al. 2022; Lin and Byrne 2022a; Chen et al. 2022). RAG-based approaches typically consist of two separate stages: retrieval and generation. In the first retrieval stage, these methods usually integrate multiple discriminative retrievers, each designed for specific purposes such as image-to-text or text-to-text retrieval. Afterward, in the second answer generation stage, these methods typically use generative multi-modal large language models (MLLM) to produce the final result. Despite achieving success in some benchmarks (Marino et al. 2019; Schwenk et al. 2022), this workflow still encounters several limitations. 1) Current methods sequentially invoke models in the pipeline for feature engineering, retrieval, and answer generation, requiring the integration of multiple heterogeneous models. 2) Moreover, these methods typically combine generative answer generators with discriminative retrievers. The disparate model architectures make it challenging for retrievers to further optimize their performance based on the feedback from the answer generator. Consequently, the research question arises: ***how can we integrate knowledge retrieval and answer generation into a homogeneous generative model?***

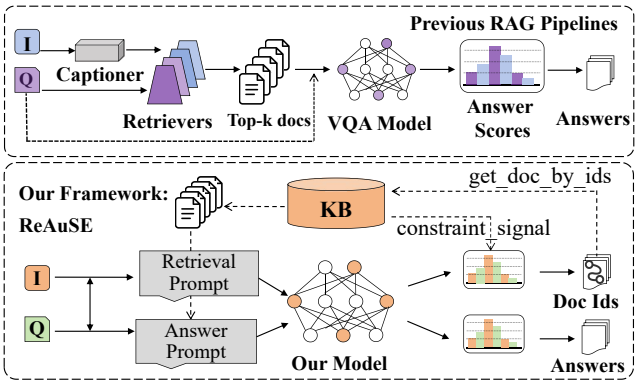

Figure 1: Comparing with the Paradigm of Previous Knowledge-Based VQA Methods.

To address the above issue, we propose ReAuSE, a novel Retrieval-augmented framework with built-in Autoregressive Search Engines for knowledge-based VQA tasks, which seamlessly integrates knowledge retrieval into the generative MLLM. ReAuSE takes advantage of the fact that MLLMs can serve as virtual knowledge ware-

*Corresponding author.

houses (Pan et al. 2024), recognizing the documents that a multi-modal query can be linked to. Therefore, ReAuSE abandons the discriminative retrieval paradigm that computing the similarity between the query and document one by one, whereas directly generates the document identifier in an autoregressive manner, where each identifier corresponds to a document within the knowledge base. We define the document identifiers as a sequence of tokens that appears at least once within a document in the knowledge base, thus enabling effective and efficient mapping to the document. Subsequently, we propose a reinforced retrieval calibration method based on relevance feedback to further enhance retrieval performance. To collect relevance preference data, we employ a MLLM as a reward model, which inputs sampled documents and questions into this model and assesses document relevance based on the VQA scores (Antol et al. 2015) of the generated answers. To align with relevance preference, we employ a direct preference optimization (DPO) algorithm (Rafailov et al. 2023) to further refine the generative retrieval model. In the answer generation stage, we input the retrieved documents one by one, and the model obtains the final prediction based on the joint probability of retrieval and answer generation.

We conduct primary experiments on two representative knowledge-based VQA benchmarks, OKVQA and A-OKVQA. The experimental results show significant improvements of 2.9%-9.6% across all metrics compared to strong baselines. Additionally, we perform knowledge retrieval experiments on three datasets to further validate the performance of the generative knowledge retrievers. Our model consistently outperforms other discriminative knowledge retrievers and the improvements become more apparent when applied to large-scale knowledge bases. This outcome illustrates our model's capability to retrieve knowledge from large-scale knowledge sources. The code will be available at https://github.com/xinwei666/ReAuSE

Related Work

Traditional Visual Question Answering (VQA) tasks (Johnson et al. 2017; Mishra et al. 2019), which focus on answering questions related to visual elements (e.g., simple counting, visual attributes), have been extensively studied. Several studies (Marino et al. 2019) have revealed that over 78% of questions can be answered by people under ten years old, indicating that traditional VQA tasks require little background knowledge to answer a vast majority of questions.

Knowledge-based VQA. To assess models' capacity to leverage world knowledge instead of relying solely on input data, knowledge-based VQA tasks have emerged, such as OKVQA (Marino et al. 2019), and A-OKVQA (Schwenk et al. 2022). OKVQA and A-OKVQA datasets pose challenges in acquiring the necessary knowledge from an outside source and performing reasoning over multi-modal contexts and knowledge. Recently, Infoseek (Chen et al. 2023d) has been proposed, featuring visual questions about detailed properties of factual knowledge in Wikipedia. The above datasets all highlight the importance of retrieving knowledge from external sources and underscore that current state-of-the-art methods still have significant room for improvement in this task.

Existing approaches have been proposed to incorporate knowledge in two ways to address knowledge-based VQA tasks. One line of research (Xenos et al. 2023a; Chen et al. 2023e; Gui et al. 2021) leverages implicit knowledge from LLMs. This approach involves converting images into text or directly feeding multi-modal contexts into LLMs (e.g. GPT-3 (Brown et al. 2020), GPT-4V (Achiam et al. 2023), etc.) to generate text that serves as augmented knowledge, but hallucinated information produced by LLMs poses risks to the overall pipeline. Another research direction (Lin et al. 2022; Hao et al. 2024a; Lin et al. 2023) aims to retrieve explicit knowledge from structured or unstructured KB. This approach, known as retrieval augmentation, often uses off-the-shelf tools to generate visual tags and captions, thereby boosting the performance of knowledge retrievers. Several studies (Gao et al. 2022; Hu et al. 2023b) have tried to combine both ways by simply using the results of LLMs and retrievers but led to limited improvements over baselines.

Knowledge Retrieval. As a crucial component of retrieval-augmented approaches, knowledge retrievers face challenges in handling multi-modal queries (Luo et al. 2021, 2023; Shen et al. 2023a). Several methods (Lin and Byrne 2022b,a; Gao et al. 2022), which employ separate text-to-text and image-to-text retrievers, struggle to capture cross-modal interactions. To bridge this gap, Reviz (Luo et al. 2023) leverages visual-language models to unify the encoding of image and text queries, and FMLR (Lin et al. 2023) proposes a fine-grained late-interaction framework to fuse cross-modal features at the token level. PreFLMR (Lin et al. 2024) explores scaling laws for knowledge retrieval based on the FLMR model. Although these methods achieve improvements over previous approaches, they require training on large-scale datasets containing millions of image-text pairs, which incurs high computational costs.

Recently, some studies (Bevilacqua et al. 2022; Ziems et al. 2023; Li et al. 2023, 2024a; Long et al. 2024b; Jain, Soares, and Kwiatkowski 2024) have introduced generative pipelines in information retrieval tasks, instead of discriminative retrievers. These methods (Tay et al. 2022) are based on the assumption that all documents are memorized by generative language models, and the language model directly generates the identifiers of relevant documents based on the query. While prior research (Li et al. 2024b; Long et al. 2024a) has investigated generative retrieval for multi-modal tasks, such methods have demonstrated only marginal gains over traditional methods when applied to general tasks. Different from them, we are the first work to seamlessly integrate generative retrieval and retrieval-augmented VQA tasks, and use the feedback from the QA module to enhance the retrieval performance, thereby achieving better retrieval and QA results simultaneously.

Methodology

We introduce ReAuSE, a **Re**trieval-**Au**gmented framework utilizing built-in **Au**toregressive **S**earch **E**ngines tailored for

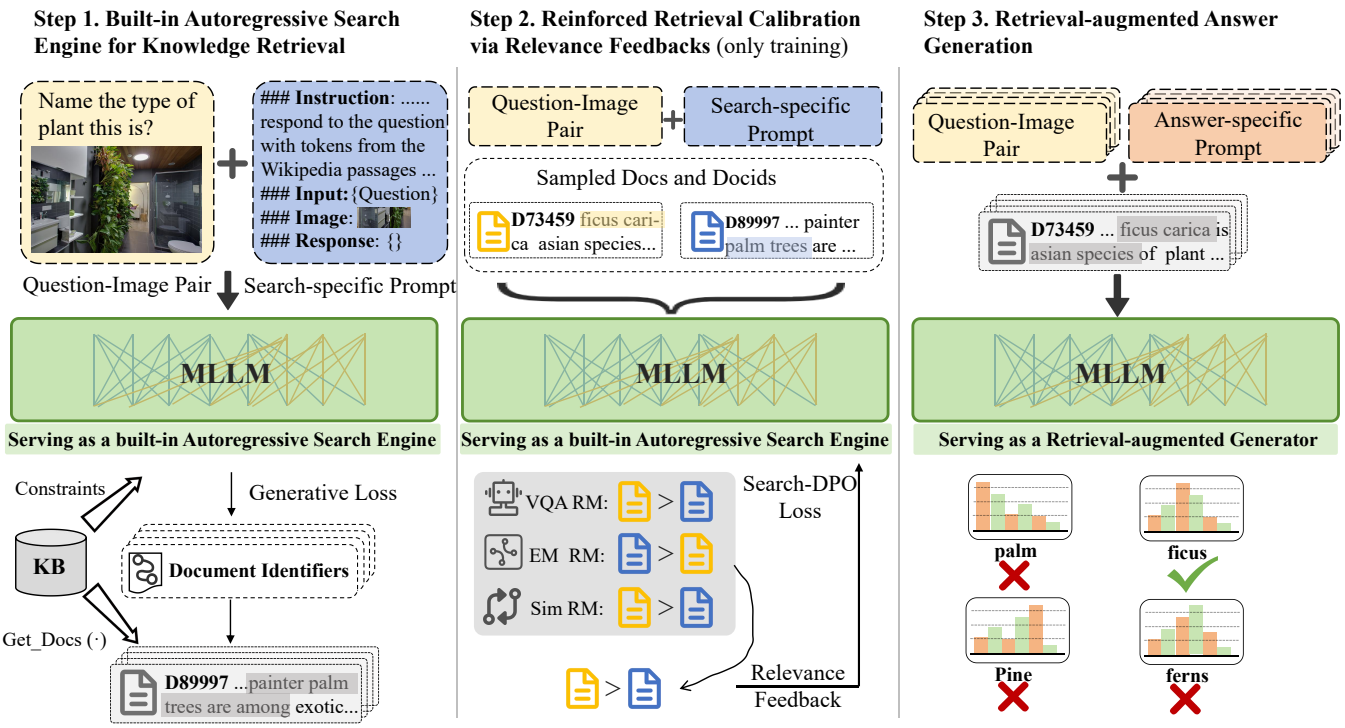

Figure 2: The architecture of ReAuSE. ReAuSE contains three components: Built-in Autoregressive Search Engine for knowledge retrieval, Reinforced Retrieval Calibration via Relevance Feedback to align retrievers with relevance preferences, and Retrieval-Augmented Generation for answer prediction.

knowledge-based VQA tasks. ReAuSE is designed as a unified model to facilitate both effective knowledge retrieval and question-answering tasks.

Problem Formulation

Formally, let $\mathcal{D} = \{D_1, ..., D_k\}$ denotes a knowledge base used for the knowledge-based VQA task, $D_i = \{d_1, ..., d_{|D|}\}$ denotes a document with its title and textual contexts, and $R_i = \{r_1, r_2, ..., r_{|R|}\}$ denotes an identifier of the document D_i. Given a multi-modal query X, the generative knowledge retrieval can be formulated as a Seq2Seq task, as Eq. 1,

$$\mathcal{P}(R_i|X) = \prod_{j=1} \mathcal{P}(r_j|\boldsymbol{r}_{<j}, X, \Theta). \quad (1)$$

where $\mathcal{P}$ denotes the standard auto-regressive language modeling probability and $\boldsymbol{\Theta}$ are the paramters of our model. During inference, the model employs a constrained strategy to guide the decoder in generating valid identifiers, which maintains a deterministic mapping relationship φ between identifier and document, as Eq. 2,

$$\varphi : R_i \rightarrow D_i, \text{where } D_i \in \mathcal{D}. \quad (2)$$

Finally, we obtain a subset $\hat{D} = \{D_1, ..., D_{|K|}\}$ from $\mathcal{D}$ to improve answer generation. The overall likelihood of generating the answer Y is given by Eq. 3,

$$\mathcal{P}(Y|X) = \sum_{D_i \in \hat{D}} \underbrace{\mathcal{P}(R_i|X)}_{retrieval} \cdot \underbrace{\mathcal{P}(Y|X, D_i)}_{generation}. \quad (3)$$

Built-in Autoregressive Search Engines

We introduce a novel autoregressive search engine for knowledge-based VQA tasks to facilitate retrieval from external knowledge bases. The autoregressive search engine leverages a generative architecture similar to that of common multimodal large language models, instead of discriminative models, enabling its seamless integration and functioning as a built-in module.

Given a multi-modal input $X = \{Q, V\}$, the autoregressive search engine aims to generate the relevant identifier directly in a seq2seq manner as Eq. 1. For example, Fig. 2 shows how our model generates the corresponding identifier for a document related to the "palm tree" based on the input image and question. To achieve such a generative retriever, we mainly elaborate on the three aspects as follows:

Document Identifier. Based on the assumption in (Pan et al. 2024) that large language models are aware of the content within each document, we define any document's identifier as subsequences that appear only in that specific document. Unlike the one-to-one relationship in DSI (Tay et al. 2022), We assign more than one identifier to each document, as long as these identifiers are unique for this document. Consequently, our model does not require additional memory steps as in existing studies (Tay et al. 2022; Li et al. 2024b) to associate documents with identifiers.

Supervised Fine-tuning teaches our model to generate relevant identifiers based on the autoregressive probability for each given multi-modal query. To sample the most rele-

vant sub-sequences from the given ground-truth document as identifiers, we employ a large language model (Touvron et al. 2023) as an extractive summarizer, which uses a fixed-length original text to answer a given question. Later, we filter the obtained set of identifiers and select the identifier containing the most answer keywords as the target identifier. Note that our model is model-agnostic, allowing it to be applied to any generative multi-modal large language model. The generative loss function can be formalized as maximizing the likelihood of the target identifier using the teacher forcing strategy, as Eq. 4.

$$\mathcal{L}_{retrieval} = \sum_{j=1} \log \mathcal{P}(r_j | \boldsymbol{r}_{<j}, X). \tag{4}$$

To avoid overfitting and catastrophic forgetting, we freeze all the parameters of the MLLM and adopt the Low-Rank Adaptation (LoRA) method (Hu et al. 2021) to efficiently fine-tune our model, with only the parameters of LoRA being updated.

Constrained Decoding and FM-Index. A valid identifier is defined as a generated sequence that appears at least once within a document in the knowledge base, ensuring that each generated identifier can be directly linked to a specific document. To help the model generate valid identifiers during inference, we implement a beam decoding strategy constrained by knowledge bases.

Specifically, we use the previously generated sequence $R_i^{t-1} = \{r_1, ..., r_{t-1}\}$ as the prefix condition to search for all matching strings in the knowledge base. We then extract the subsequent tokens from these strings to form a feasible token set $\mathcal{S}$. The model's next token, r_t, is restricted to selection from $\mathcal{S}$, guaranteeing that all generated sequences exist within the knowledge base. To support fast substring search, we utilize an FM-Index database (Ferragina and Manzini 2000; Bevilacqua et al. 2022) to store the knowledge base. FM-Index is an efficient indexing structure tailored for substring search. The time complexity for obtaining the next allowed token is nearly $\mathcal{O}(V)$, where V is the vocabulary size, independent of the size of the knowledge base.

Reinforced Retrieval Calibration via Relevance Feedback

Despite teaching our model through supervised fine-tuning to generate relevant document identifiers based on user queries, the retrieved documents exhibit varying degrees of relevance. Even when documents are provided, the QA model may struggle to provide accurate responses. Optimally, the generative retriever should retrieve documents that: (1) strongly correlate with the multi-modal query, and (2) minimize extraneous content. Consequently, it is essential to further improve retrieval performance through feedback from the QA model.

As the first step towards this goal, we sample a set of identifiers $\{R_1, ..., R_k\}$ for each X using the generative retriever π_{sft} that has been supervised fine-tuned. Then, we score the collected samples by evaluating their relevance from three aspects:

- **Contributions to VQA performance.** A document is deemed relevant if a model can produce the correct answer using it. To evaluate this relevance, we employ an MLLM that has not been fine-tuned on downstream data as the reward model, with the VQA score serving as the reward value $v_{vqa} \in [0, 1]$.
- **Keyword Hit Count.** If an identifier includes keywords from the answer set, it is likely to be relevant. To quantify this relevance, we employ an exact matching function as the reward function, with matching signals serving as the reward values $v_{hit} \in \{0, 1\}$.
- **Semantic Similarity.** Higher semantic similarity between an identifier and a document indicates that the identifier better represents the document's semantics, thereby suggesting a lower presence of irrelevant content within the document. To measure this relevance, we use the BERT model to calculate the cosine similarity between identifiers and documents as the reward values $v_{sim} \in [0, 1]$.

The overall reward can be obtained by taking a weighted sum of the scores from different aspects. Then, we build a triplet $< X, R^+, R^- >$ for each X by treating the identifiers with the highest/lowest reward as positive/negative samples, respectively. Using the triplets reflecting the QA model's preference, the retriever can be further aligned by preference-based reinforcement learning. As one of the typical methods, direct preference optimization (DPO) (Rafailov et al. 2023) is widely used for its efficiency and effectiveness. Therefore, we employ the DPO loss to further optimize our autoregressive knowledge retriever as Eq. 5,

$$\mathcal{L}_{dpo} = -\log\sigma\left(\beta \log \frac{\pi_\Theta(R^+|X)\pi_{sft}(R^-|X)}{\pi_{sft}(R^+|X)\pi_\Theta(R^-|X)}\right). \tag{5}$$

where π_{sft} is the original model used as reference, and π_Θ is the model being optimized. As before, we only update the parameters of LoRA.

Answer Generation

Utilizing built-in autoregressive knowledge retrievers, we extract the top-K relevant documents from extensive knowledge bases to serve as external knowledge. For our answer generation model, we employ a model architecture homologous to that of the retrieval module. As illustrated in Fig. 2, we construct a prompt template, filling the slots with the image, question, and each retrieved document. The multi-modal contexts are then fed into the model, and the training loss of the answer generation follows that of the generative retrieval model, as Eq. 6,

$$\mathcal{L}_{gen} = \sum_{j=1} \log \mathcal{P}(y_j | \boldsymbol{y}_{<j}, X, D_i). \tag{6}$$

where y_j denotes the $j-th$ token of the ground-truth answer Y. As before, we freeze all the parameters of the MLLM, but introduce another LoRA, and only update the parameters of this new LoRA.

$$\begin{aligned} \hat{Y}, \hat{D} &= \underset{Y, D_i}{\operatorname{argmax}} \mathcal{P}(Y, D_i | X) \\ &= \underset{Y, D_i}{\operatorname{argmax}} \mathcal{P}(Y | X, D_i) \cdot \mathcal{P}(R_i | X). \end{aligned} \tag{7}$$

During inference, We use the same MLLM and parameters for both the retrieval and answer generation stages, except for the two LoRA adapters. After retrieving the relevant document set, we switched to the LoRA adapter for answer generation, and obtain the final prediction through the joint probability of retrieval and answer generation, as Eq. 7.

Experiments

Experiment Setup

Datasets and Knowledge Bases. We focus on the knowledge-based VQA benchmarks, OKVQA (Marino et al. 2019) and A-OKVQA (Schwenk et al. 2022). Previous work provided two retrieval corpora, GS112K (Luo et al. 2021) and Wiki21M (Karpukhin et al. 2020), for the OKVQA dataset. GS112K contains 112K passages collected through Google Search, while Wiki21M is a subset of Wikipedia, containing 21M Wikipedia entries. Moreover, we also conduct retrieval experiments on these two corpora and introduce a new information-seeking dataset, InfoSeek (Chen et al. 2023d), to evaluate the model's retrieval performance. Since InfoSeek's KB is not publicly available, we use the KB provided by PreFLMR (Lin et al. 2024) and follow the same experimental setup.

Evaluation Metrics. We strictly follow the settings of the original papers, using the corresponding metrics for each dataset. For the OKVQA dataset and the "direct answer" setting of the A-OKVQA dataset, we use the VQA score to evaluate the model's performance. For the "multi-choice" setting of the A-OKVQA dataset, we use accuracy for evaluation. To evaluate the performance of knowledge retrieval, we use the Pseudo-relevance Recall@K (PRR@K) (Luo et al. 2021), consistent with the baselines.

Baselines. We adopt several baseline methods for comparison, categorized as follows: 1) multi-modal large language models: LLaVA-13B (Liu et al. 2023), PALM-E-562B (Chen et al. 2023c), and GPT-4V (Achiam et al. 2023). 2) knowledge-enhanced methods via GPT-3/4 APIs: Prophet (Shao et al. 2023), Promptcap (Hu et al. 2023a), FillingGap (Wang et al. 2023) and REVIVE (Lin et al. 2022). 3) retrieval-augmented methods: TwO (Si et al. 2023), ReVeaL (Hu et al. 2023b), GeMKR (Long et al. 2024a), and FLMR (Lin et al. 2023). For the A-OKVQA dataset, we also add the advanced GPV-2 (Schwenk et al. 2022), SimVQA (Xenos et al. 2023b), Cola-FT(11B+3B) (Chen et al. 2023b) and CKR-VQA (Hao et al. 2024a) as baselines.

Implementation Details. Our framework is model-agnostic. In our main experiments, we utilize MiniGPT4-v2-7B as the base model, which employ ViT-L/14 from pre-trained CLIP as the image encoder and LLaMa-v2-7B (Touvron et al. 2023) as the text encoder. We freeze all parameters of the MLLM, allowing updates only to the LoRA parameters. **We use the same MLLM in the three stages but apply two sets of LoRA parameters to optimize the model respectively: one for retrieval and alignment, and the other for answer generation.** Our model is implemented in PyTorch, utilizing version 0.3.0 of the PEFT library, which supports efficient switching between two LoRA adapters during inference. Similar to baselines, we use image captions as features to enhance the model's performance. Each training stage is performed on four NVIDIA A6000 48G GPUs and completed within three hours.

Model	PRR@K	Score
Multi-modal Large Language Models		
LLaVA-13B (Liu et al. 2023)	-	61.9
Minigpt4-v2-7B (Chen et al. 2023a)	-	57.8
Minigpt4-v2-7B (FT) (Chen et al. 2023a)	-	61.9
PaLM-E-562B (Driess et al. 2023)	-	66.1
GPT-4V (Achiam et al. 2023)	-	64.3
Knowledge-enhanced Methods via GPT-3/4v APIs		
ReVIVE (Lin et al. 2022)	-	58.0
Prophet (Shao et al. 2023)	-	61.1
Promptcap (Hu et al. 2023a)	-	60.4
FillingGap (Wang et al. 2023)	-	61.3
MM-Reasoner (Khademi et al. 2023)	-	60.8
Retrieval-augmented Generation Methods		
TRiG (Gao et al. 2022)	45.8	50.5
RA-VQA (Lin and Byrne 2022a)	82.8	54.5
TwO (Si et al. 2023)	-	56.7
ReVeaL (Hu et al. 2023b)	-	59.1
FLMR (Lin et al. 2023)	89.3	62.1
FLMR (Lin et al. 2023) $*$	88.3	62.7
KSVQA (Hao et al. 2024b)	-	62.8
GeMKR (Long et al. 2024a) $*$	78.6	61.8
ReAuSE (Ours)	**92.6**	**65.7**

Table 1: Performance on the OKVQA benchmark. PPR@K applies only to RAG baselines; "-" denotes inapplicability or unavailable results. "$*$" indicates the results we reproduced using the official code and the same answer generator as our model.[1]

Main Results

We compare our ReAuSE with the aforementioned baselines for knowledge-based VQA tasks in Tab. 1 and Tab. 2. The experimental results illustrate that ReAuSE achieves significant improvements over the competitive baselines on the challenging OKVQA and A-OKVQA datasets.

From Tab. 1, we can observe that ReAuSE outperforms the competitive baseline FLMR on both retrieval and VQA metrics, which consistently demonstrates the effectiveness of our method in integrating both knowledge retrieval and answer generation into a unified multi-modal large language model framework. ReAuSE achieves an advanced VQA score on OKVQA when compared to models with similar parameter scales, surpassing the previous best retrieval-augmented method by more than 2.9% and outperforming methods that use LLM-APIs for knowledge enhancement by 4.6%. Moreover, our method exceeds GPT-4V by 1.45% in VQA score. Even compared with the closed-source PALM-E-562B, which is over 80 times larger than ours, our method is only 0.5% behind.

[1]We first use officially released checkpoints to obtain retrieved documents and then feed these documents into the answer generator (fine-tuned Minigpt4-v2) to acquire the corresponding answers.

Models	Multi-Choice		Direct-Answer	
	val	test	val	test
LLaVA-1.5-7B	77.1	74.5	63.7	58.6
InstructBLIP-7B(FT)	73.0	71.1	62.4	58.7
Minigpt4-v2-7B(FT)	-	-	61.3	-
GPV-2	60.3	53.7	48.6	40.7
PromptCap	73.2	73.1	56.3	59.6
Prophet	76.4	73.6	58.2	55.7
FillingGap	-	-	59.8	-
SimVQA	-	-	58.6	57.5
REVEAL	-	-	52.2	-
Cola-FT	78.1	76.7	-	-
CKR-VQA	76.2	75.4	58.1	60.1
ReAuSE (Ours)	**85.0**	**80.3**	**67.7**	**65.8**

Table 2: Performance on the A-OK-VQA benchmark.

Ablation Setting	PRRecall@5	Score
Full Model (Ours)	**92.6**	**65.7**
w/o Search Engines	-	61.9
w/o Fine-tuning Search Engine	33.1	61.6
w/o Constrained Decoding	-	63.2
w/o Retrieval Calibration	88.7	62.5
w/o VQA Reward Model	91.0	63.3
w/o EM Reward Func.	89.9	64.5
w/o Sim. Reward Model	91.7	65.3

Table 3: Ablation Studies. *w/o* denotes "without".

The OKVQA benchmark poses a challenging issue of retrieving relevant knowledge from extensive knowledge bases or directly generating useful information about multimodal contexts. Despite using GPT-3 or GPT-4V to acquire knowledge or directly adopting GPT-3 as the backbone, MM-Reasoner and FillingGap fail to achieve obvious improvements compared to retrieval-augmented methods. In contrast, retrieval-augmented methods, such as FLMR and KSVQA, achieve better VQA performance by incorporating manually designed feature engineering and integrating multiple retrievers and selectors.

From Tab. 2, ReAuSE demonstrates more significant performance improvements on A-OKVQA, with accuracy and VQA scores increasing by 4.9% to 9.6% compared to baselines of similar parameter scales[2]. Our approach demonstrates consistent improvements, which can be attributed to two key factors. First, we leverage large language models as virtual knowledge bases by replacing traditional discriminative pipelines with generative retrievers. Second, we implement reinforced retrieval calibration to align the search engine with the answer generator, enabling the retriever to incorporate relevance feedback for refinement, thereby yielding more relevant results. In the following sections, we will examine the performance of the autoregressive search engine and analyze the impact of search results on the answer generation process.

[2]See the submission at: https://leaderboard.allenai.org/a-okvqa/submission/cqp56m03c8g0k0quidj0

Ablation Study

We conduct a series of ablation studies by gradually removing each module of our framework and the corresponding results are presented in Tab. 3.

To evaluate the impact of retrieval augmentation, we first remove the built-in autoregressive search engine, using the MLLM as an answer generator without access to external knowledge. This operation results in a 3.8% decrease in the VQA score, indicating that external knowledge retrieval is crucial for knowledge-based VQA tasks. Next, if we do not supervised fine-tune the MLLMs, it cannot effectively serve as a generative search engine to retrieve knowledge from the KB. Moreover, we disable the constrained decoding strategy, allowing the MLLM to generate image-related knowledge without restrictions. However, since this freely generated content cannot be linked to the document in the KB, it is used directly as external knowledge to support the answer generation process. This approach leads to a 2.5% decrease in the VQA score, likely due to the MLLM producing erroneous or hallucinated information, which results in inaccurate outputs from the answer generator.

To evaluate the effectiveness of the Reinforced Retrieval Calibration (RRC) module, we employ the generative search engine after supervised fine-tuning but remove the reinforced calibration module. We observe a 3.9% decrease in retrieval performance, which is slightly below that of the strongest baseline, FLMR. This suggests that the autoregressive retriever can be further optimized through the RRC module by leveraging relevance feedback from reward models. Furthermore, we disable each reward model to assess its effectiveness. We find that the VQA reward model enables the generative retriever to retrieve documents that align with the answer generator's preferences, thereby improving VQA performance. Conversely, the EM reward model ensures that the generated identifiers include answer keywords, leading to enhanced retrieval performance.

Effects of Retrieval Performance

To assess our model's capability in retrieving knowledge from large-scale knowledge bases, we conduct experiments on the OKVQA dataset using two retrieval corpora: Google Search (GS112K) (Luo et al. 2021) and Wikipedia (Wiki21M) (Karpukhin et al. 2020), with knowledge bases ranging in size from 112K to 21M documents. Additionally, we introduce a new dataset, Infoseek (Chen et al. 2023d), consisting of 100K documents, which poses challenges for visual entity retrieval. As shown in Tab. 4, our proposed approach consistently outperforms the leading state-of-the-art baselines FLMR and Pre-FLMR across all evaluated metrics. Specifically, our model outperforms FLMR by 3.3% in PRRecall@5 on the GS112K corpus. This improvement explains why our answer generation model surpasses the FLMR model by 3.6% in the VQA score, as shown in Tab. 1. Moreover, our method outperforms FLMR by 10.6% on the Infoseek dataset and surpasses PreFLMR by 1.7%, indicating the effectiveness of ReAuSE in handling visual entity retrieval tasks.

We observe a significant performance drop of over 20% in the FLMR model when applied to the Wiki21M cor-

#	Retrievers	OKVQA- GS112K		OKVQA- WK21M		InfoSeek-100K
		PRRecall@5	PRRecall@10	PRRecall@5	PRRecall@10	PRRecall@5
1	DPR (Karpukhin et al. 2020)	83.4	90.3	66.9	76.4	-
2	RA-VQA (Lin and Byrne 2022a)	82.8	89.0	-	-	-
3	ReViz-ICT (Luo et al. 2023)	73.4	83.2	61.9	72.6	-
4	GeMKR (Long et al. 2024a)	78.6	86.2	70.8	79.1	48.9
5	FLMR (Lin et al. 2023)	89.3	94.0	68.1	78.0	47.1
6	Pre-FLMR (Lin et al. 2024)	-	-	68.6	-	57.8
7	**ReAuSE (Ours)**	**92.6**	**95.8**	**88.0**	**91.3**	**59.5**

Table 4: Retrieval Performance on Three Retrieval Corpora.

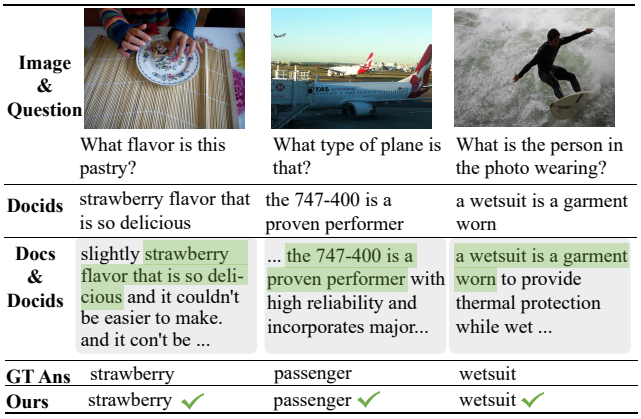

Figure 3: Case Studies. The highlighted text represents the document identifier generated by our model.

Models	TopK	GS112K	Wiki21M
DPR	5	370.4ms	518.4ms
FLMR	5	758.4ms	-
ReAuSE (Ours)	5	751.3ms	962.1ms
ReAuSE (Ours)	10	1023.3ms	1273.2.1ms

Table 5: The Retrieval Time in the Inference Stage.

pus, while our model exhibits only a 4.6% decrease. This indicates that our model demonstrates stronger generalization capabilities for retrieving from large-scale corpora. This can be attributed to the advantages of generative search engines, which generate document identifiers through token-level search rather than relying on one-to-one matches at the document level (i.e., document-level search). Although the number of documents increases, the size of the token set (i.e., vocabulary) does not expand proportionally. Consequently, generative search engines are less affected by the scale of the knowledge base, whereas the performance of discriminative methods degrades as the corpus size increases.

Efficiency

ReAuSE requires fewer resources than other retrieval-augmented baselines. ReAuSE unifies three stages into a single MLLM, whereas other baselines require an MLLM for generation and at least one traditional retriever for retrieval. ReAuSE uses LoRA fine-tuning, requiring 9K (OKVQA) to 17K (A-OKVQA) training data to train 0.49% of the parameters, with both SFT and RRC completing within 3 hours across 4 GPUs. In contrast, traditional retrievers such as PreFLMR and ReViz-ICT necessitate additional millions of data for full-scale fine-tuning.

We record the inference time of ReAuSE, FLMR, and the most efficient baseline, DPR to provide a qualitative result. As shown in Tab. 5, ReAuSE has comparable efficiency to traditional retrievers. ReAuSE generates top-K document identifiers with l tokens for each query through a l-steps decoding ($l = 10$). In contrast, for each query, traditional retrievers need to calculate the similarity with many documents. When TopK=5, our model is only 440 milliseconds slower than DPR. Considering the nearly 20% performance improvement, we argue that such latency is acceptable.

Case Study

As illustrated in Fig. 3, ReAuSE accurately generates the correct answers for all three samples. ReAuSE directly generates document identifiers associated with the image-text pairs using its built-in search engine. Each document identifier is a sequence of tokens representing a document, and it can be linked to a corresponding document that potentially contains information to answer the given question. What's more, we observe that all generated document identifiers contain answer keywords, suggesting that generated document identifiers are highly relevant to the question.

Conclusion

In this paper, we introduce ReAuSE, a novel KBVQA approach by integrating knowledge retrieval and generation within a unified generative multi-modal large language model (MLLM) framework. Extensive experimental results have shown that ReAuSE consistently outperforms existing methods, achieving significant improvements across various evaluation metrics on two benchmarks. Future work will focus on extending the application of ReAuSE to domains such as biomedicine and education (Gao et al. 2021).

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2022ZD0160603), the NSFC (No. 62406161), CPSF (No. 2023M741950),

and the Postdoctoral Fellowship Program of CPSF (No. GZB20230347).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*, 2425–2433.

Bevilacqua, M.; Ottaviano, G.; Lewis, P. S. H.; Yih, S.; Riedel, S.; and Petroni, F. 2022. Autoregressive Search Engines: Generating Substrings as Document Identifiers.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners.

Chan, C.-M.; Xu, C.; Yuan, R.; Luo, H.; Xue, W.; Guo, Y.; and Fu, J. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024. Benchmarking Large Language Models in Retrieval-Augmented Generation. In *AAAI*, 17754–17762.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chen, L.; Li, B.; Shen, S.; Yang, J.; Li, C.; Keutzer, K.; Darrell, T.; and Liu, Z. 2023b. Large Language Models are Visual Reasoning Coordinators.

Chen, W.; Hu, H.; Chen, X.; Verga, P.; and Cohen, W. W. 2022. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *EMNLP*, 5558–5570.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2023c. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *The Eleventh International Conference on Learning Representations*.

Chen, Y.; Hu, H.; Luan, Y.; Sun, H.; Changpinyo, S.; Ritter, A.; and Chang, M. 2023d. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *EMNLP*, 14948–14968.

Chen, Z.; Zhou, Q.; Shen, Y.; Hong, Y.; Zhang, H.; and Gan, C. 2023e. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Ferragina, P.; and Manzini, G. 2000. Opportunistic Data Structures with Applications. In *FOCS*, 390–398.

Gao, F.; Ping, Q.; Thattai, G.; Reganti, A. N.; Wu, Y. N.; and Natarajan, P. 2022. Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering. In *CVPR*, 5057–5067.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C. H.; Wang, X.; and Li, H. 2019. Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. In *CVPR*, 6639–6648.

Gao, W.; Liu, Q.; Huang, Z.; Yin, Y.; Bi, H.; Wang, M.-C.; Ma, J.; Wang, S.; and Su, Y. 2021. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 501–510.

Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.; Bisk, Y.; and Gao, J. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Hao, D.; Jia, J.; Guo, L.; Wang, Q.; Yang, T.; Li, Y.; Cheng, Y.; Wang, B.; Chen, Q.; Li, H.; and Liu, J. 2024a. Knowledge Condensation and Reasoning for Knowledge-based VQA. *CoRR*, abs/2403.10037.

Hao, D.; Wang, Q.; Guo, L.; Jiang, J.; and Liu, J. 2024b. Boter: Bootstrapping Knowledge Selection and Question Answering for Knowledge-based VQA. *arXiv preprint arXiv:2404.13947*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, R.; Andreas, J.; Rohrbach, M.; Darrell, T.; and Saenko, K. 2017. Learning to Reason: End-to-End Module Networks for Visual Question Answering. In *ICCV*, 804–813.

Hu, Y.; Hua, H.; Yang, Z.; Shi, W.; Smith, N. A.; and Luo, J. 2023a. PromptCap: Prompt-Guided Image Captioning for VQA with GPT-3. In *ICCV*, 2951–2963.

Hu, Z.; Iscen, A.; Sun, C.; Wang, Z.; Chang, K.; Sun, Y.; Schmid, C.; Ross, D. A.; and Fathi, A. 2023b. Reveal: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory. In *CVPR*, 23369–23379.

Jain, P.; Soares, L. B.; and Kwiatkowski, T. 2024. From RAG to Riches: Retrieval Interlaced with Sequence Generation. In *EMNLP*, 8887–8904.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. B. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*, 1988–1997.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Khademi, M.; Yang, Z.; Frujeri, F.; and Zhu, C. 2023. MM-Reasoner: A Multi-Modal Knowledge-Aware Framework for Knowledge-Based Visual Question Answering. In *EMNLP (Findings)*, 6571–6581.

Li, Y.; Lin, X.; Wang, W.; Feng, F.; Pang, L.; Li, W.; Nie, L.; He, X.; and Chua, T.-S. 2024a. A survey of generative search

and recommendation in the era of large language models. *arXiv preprint arXiv:2404.16924*.

Li, Y.; Wang, W.; Qu, L.; Nie, L.; Li, W.; and Chua, T.-S. 2024b. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. *arXiv preprint arXiv:2402.10805*.

Li, Y.; Yang, N.; Wang, L.; Wei, F.; and Li, W. 2023. Multiview identifiers enhanced generative retrieval. *arXiv preprint arXiv:2305.16675*.

Lin, W.; and Byrne, B. 2022a. Retrieval Augmented Visual Question Answering with Outside Knowledge. In *EMNLP*, 11238–11254.

Lin, W.; and Byrne, B. 2022b. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Lin, W.; Chen, J.; Mei, J.; Coca, A.; and Byrne, B. 2023. Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering. In *NeurIPS*.

Lin, W.; Mei, J.; Chen, J.; and Byrne, B. 2024. PreFLMR: Scaling Up Fine-Grained Late-Interaction Multi-modal Retrievers. *arXiv preprint arXiv:2402.08327*.

Lin, Y.; Xie, Y.; Chen, D.; Xu, Y.; Zhu, C.; and Yuan, L. 2022. REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering. In *NeurIPS*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Long, X.; Zeng, J.; Meng, F.; Ma, Z.; Zhang, K.; Zhou, B.; and Zhou, J. 2024a. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18733–18741.

Long, X.; Zeng, J.; Meng, F.; Zhou, J.; and Zhou, B. 2024b. Trust in Internal or External Knowledge? Generative Multi-Modal Entity Linking with Knowledge Retriever. In *ACL (Findings)*, 7559–7569.

Luo, M.; Fang, Z.; Gokhale, T.; Yang, Y.; and Baral, C. 2023. End-to-end Knowledge Retrieval with Multi-modal Queries. In *ACL (1)*, 8573–8589.

Luo, M.; Zeng, Y.; Banerjee, P.; and Baral, C. 2021. Weakly-Supervised Visual-Retriever-Reader for Knowledge-based Question Answering. In *EMNLP (1)*, 6417–6431.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*, 3195–3204.

Mascharka, D.; Tran, P.; Soklaski, R.; and Majumdar, A. 2018. Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning. In *CVPR*, 4942–4950.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. OCR-VQA: Visual Question Answering by Reading Text in Images. In *ICDAR*, 947–952.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7): 3580–3599.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *ECCV (8)*, volume 13668 of *Lecture Notes in Computer Science*, 146–162.

Shah, S.; Mishra, A.; Yadati, N.; and Talukdar, P. P. 2019. KVQA: Knowledge-Aware Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 8876–8884.

Shao, Z.; Yu, Z.; Wang, M.; and Yu, J. 2023. Prompting Large Language Models with Answer Heuristics for Knowledge-Based Visual Question Answering. In *CVPR*, 14974–14983.

Shen, F.; Shu, X.; Du, X.; and Tang, J. 2023a. Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval. In *Proceedings of the 31th ACM International Conference on Multimedia*.

Shen, F.; Xie, Y.; Zhu, J.; Zhu, X.; and Zeng, H. 2023b. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing*, 32: 1039–1051.

Si, Q.; Mo, Y.; Lin, Z.; Ji, H.; and Wang, W. 2023. Combo of Thinking and Observing for Outside-Knowledge VQA. In *ACL (1)*, 10959–10975.

Sun, H.; Xu, L.; Jin, S.; Luo, P.; Qian, C.; and Liu, W. 2024. PROGRAM: PROtotype GRAph Model based Pseudo-Label Learning for Test-Time Adaptation. In *The Twelfth International Conference on Learning Representations*.

Tay, Y.; Tran, V.; Dehghani, M.; Ni, J.; Bahri, D.; Mehta, H.; Qin, Z.; Hui, K.; Zhao, Z.; Gupta, J. P.; Schuster, T.; Cohen, W. W.; and Metzler, D. 2022. Transformer Memory as a Differentiable Search Index. In *NeurIPS*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, Z.; Chen, C.; Li, P.; and Liu, Y. 2023. Filling the Image Information Gap for VQA: Prompting Large Language Models to Proactively Ask Questions. 2874–2890.

Xenos, A.; Stafylakis, T.; Patras, I.; and Tzimiropoulos, G. 2023a. A Simple Baseline for Knowledge-Based Visual Question Answering. In *EMNLP*, 14871–14877.

Xenos, A.; Stafylakis, T.; Patras, I.; and Tzimiropoulos, G. 2023b. A simple baseline for knowledge-based visual question answering. *arXiv preprint arXiv:2310.13570*.

Zhu, X.; Qi, B.; Zhang, K.; Long, X.; Lin, Z.; and Zhou, B. 2024. PaD: Program-aided Distillation Can Teach Small Models Reasoning Better than Chain-of-thought Fine-tuning. In *NAACL-HLT*, 2571–2597.

Ziems, N.; Yu, W.; Zhang, Z.; and Jiang, M. 2023. Large language models are built-in autoregressive search engines. *arXiv preprint arXiv:2305.09612*.