Wiki-LLaVA:
Hierarchical Retrieval-Augmented Generation for Multimodal LLMs

Davide Caffagni[1,*] Federico Cocchi[1,2,*] Nicholas Moratelli[1,*] Sara Sarto[1,*]
Marcella Cornia[1] Lorenzo Baraldi[1] Rita Cucchiara[1,3]
[1]University of Modena and Reggio Emilia, Italy [2]University of Pisa, Italy [3]IIT-CNR, Italy
[1]{name.surname}@unimore.it [2]{name.surname}@phd.unipi.it

Abstract

Multimodal LLMs are the natural evolution of LLMs, and enlarge their capabilities so as to work beyond the pure textual modality. As research is being carried out to design novel architectures and vision-and-language adapters, in this paper we concentrate on endowing such models with the capability of answering questions that require external knowledge. Our approach, termed Wiki-LLaVA, aims at integrating an external knowledge source of multimodal documents, which is accessed through a hierarchical retrieval pipeline. Relevant passages, using this approach, are retrieved from the external knowledge source and employed as additional context for the LLM, augmenting the effectiveness and precision of generated dialogues. We conduct extensive experiments on datasets tailored for visual question answering with external data and demonstrate the appropriateness of our approach.

1. Introduction

Recently, Large Language Models (LLMs) have demonstrated impressive performance in zero-shot textual tasks. Specifically, recent literature has devised models capable of tackling diverse tasks, as instructed by the user [6, 30, 41]. In this context, the classical approach is that of fine-tuning a model on varied tasks that are described through natural language [7, 34], thus empowering the model to assimilate externally provided instructions and facilitating robust generalization across multiple domains. Following these advancements, the computer vision community has started to investigate the extension of such models to vision-and-language contexts, thus generating Multimodal Large Language Models (MLLMs). On this line, the fusion of visual features into LLM backbones through vision-to-language adapters [1, 21, 23, 48] has induced notable performance improvements, enabling extensive generalization to vision-and-language tasks requiring elaborate visual descriptions.

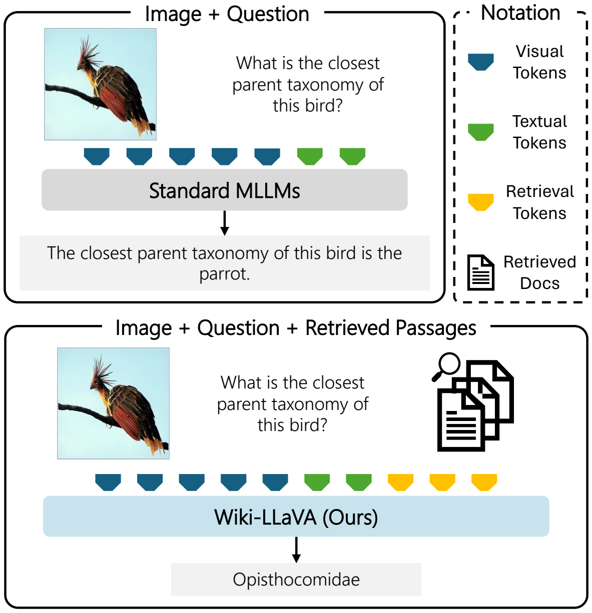


Image + Question
What is the closest parent taxonomy of this bird?
Standard MLLMs
The closest parent taxonomy of this bird is the parrot.
Notation
Visual Tokens
Textual Tokens
Retrieval Tokens
Retrieved Docs
Image + Question + Retrieved Passages
What is the closest parent taxonomy of this bird?
Wiki-LLaVA (Ours)
Opisthocomidae

Figure 1. Comparison between a standard multimodal LLM and Wiki-LLaVa. Our model integrates knowledge retrieved from an external knowledge base of documents through a hierarchical retrieval pipeline. As a result, it provides more precise answers when tasked with questions that require external knowledge.

In this context, MLLMs excel by simply including a small module (*i.e.*, an adapter) that aligns visual features with textual ones. However, despite these models being built upon LLMs trained on large-scale data, they exhibit notable limitations when confronted with highly specific user queries or when a certain degree of compositional reasoning is required to formulate the response. Moreover, certain knowledge proves itself challenging to be encoded within the parameters of an MLLM, due to the scarcity of long-tail information in the training data. In response to this challenge, different benchmarks have

[*] Equal contribution.

been recently introduced for evaluating the capabilities of MLLM to tackle queries related to external data, such as InfoSeek [5] and Encyclopedic-VQA [28]. While different works [8, 20, 21, 32] have been testing on these benchmarks, underscoring the significance of this area, none of them has developed architectures specifically designed for tackling external knowledge.

Driving from these considerations, in this paper we propose the first MLLM augmented with a retrieval module, thus shifting the focus towards teaching the model to leverage diverse information in its responses and learning to discern the relative importance of each. In particular, our model retrieves appropriate information from an external knowledge base of documents and employs a hierarchical retrieval approach to identify relevant passages. This additional knowledge is then fed to an MLLM, without changing its structure but improving its answering capabilities. To the best of our knowledge, our work represents the first MLLM to harness the retrieval capability of external sources. We assess the quality of the proposed approach by conducting extensive experiments and comparisons with respect to recent MLLMs [8, 21, 24] and by showcasing the effectiveness of our design choices. Experimental results demonstrate the advantage of retrieving from external sources and the appropriateness of our model design. Overall, we conceive our work as a first step in the direction of retrieval-augmented MLLMs, which could foster future works in the same area.

2. Related Work

Multimodal LLMs. LLMs have significantly reshaped the landscape of AI research and applications, spearheaded by notable examples like OpenAI's ChatGPT and GPT-4. These models leverage alignment techniques such as instruction tuning [30] and reinforcement learning from human feedback [39] and achieve remarkable capabilities in language understanding and reasoning. Open-source LLMs like Flan-T5 [7], Vicuna [6], LLaMA [41], and Alpaca [40] have further accelerated the advancement within the research community. This surge in the development of LLMs subsequently led to the emergence of MLLMs [3], which can combine the understating of visual inputs with natural language generation.

Early attempts of building MLLMs such as VisualGPT [4] and Frozen [42] used pre-trained language models to enhance vision-and-language models specifically for tasks like image captioning and visual question answering. This initial investigation paved the way for subsequent research in this domain, with the introduction of solutions such as Flamingo [1] or BLIP-2 [21] which allowed the integration of image features into LLMs respectively through trainable cross-attention layers directly within the LLM or Q-Former blocks that instead combine image and textual features via learnable queries. Building upon these advancements, subsequent models like FROMAGe [19], Kosmos-1 [14], and MiniGPT-4 [48] have been introduced to further refine the interplay between visual and language modalities within the LLM architecture.

Concurrently, the LLaVA family of models [23–25] introduced the usage of instruction tuning in the multimodal domain, by training on a curated dataset collected with GPT-4. This strategy is now among the most promising recipes for building MLLMs.

Retrieval-augmented language models. In recent years, retrieval-augmentation has been applied to language models by expanding their input space with relevant text passages extracted from external sources [10] or eventually retrieved directly from the web [29]. These techniques have demonstrated large improvements in knowledge-intensive tasks and significant savings in terms of model size.

Traditionally, the integration of external knowledge into textual generation has been confined to the initial stages. Different solutions [17] proposed to adaptively retrieve passages for generation on top of a proprietary LLM. Some works [10], instead, focused on capturing knowledge in a more modular and interpretable way, by augmenting the language model pre-training with a latent knowledge retriever. This allows the model to retrieve and attend documents taken from a large corpus such as Wikipedia.

While much attention has been directed towards textual augmentation, similar research efforts have recently been dedicated in the context of vision-and-language tasks [2, 13, 31, 37]. Following this direction, the work presented in [13] proposed a retrieval-augmented visual-language model that encodes world knowledge into a large-scale memory. Other approaches [35, 36] also apply retrieval to specific downstream tasks such as image captioning. Differently from all the aforementioned approaches, our work is the first to apply retrieval-augmentation to MLLMs. We do this by applying a hierarchical retrieval strategy on top of a knowledge base made of multimodal documents.

Knowledge-based visual question answering. Recently, the emergence of new benchmarks like Encyclopedic-VQA [28] and InfoSeek [5] has raised the difficulty of standard knowledge-based VQA [16, 27, 38] with questions that require intensive knowledge about specific entities, such that even LLM-based models perform poorly without retrieving information from external sources. Often, contrastive image-text encoders are employed to retrieve the target entity given the query image [44, 46]. Then, the entity name is used as a key to access an external knowledge base, which is typically composed of several text passages that encompass the correct answer. In this work, we design a hierarchical retrieval scheme based on CLIP [33] and the Contriever model [15] to extrapolate relevant passages, and we feed them to an MLLM to help the answer generation.

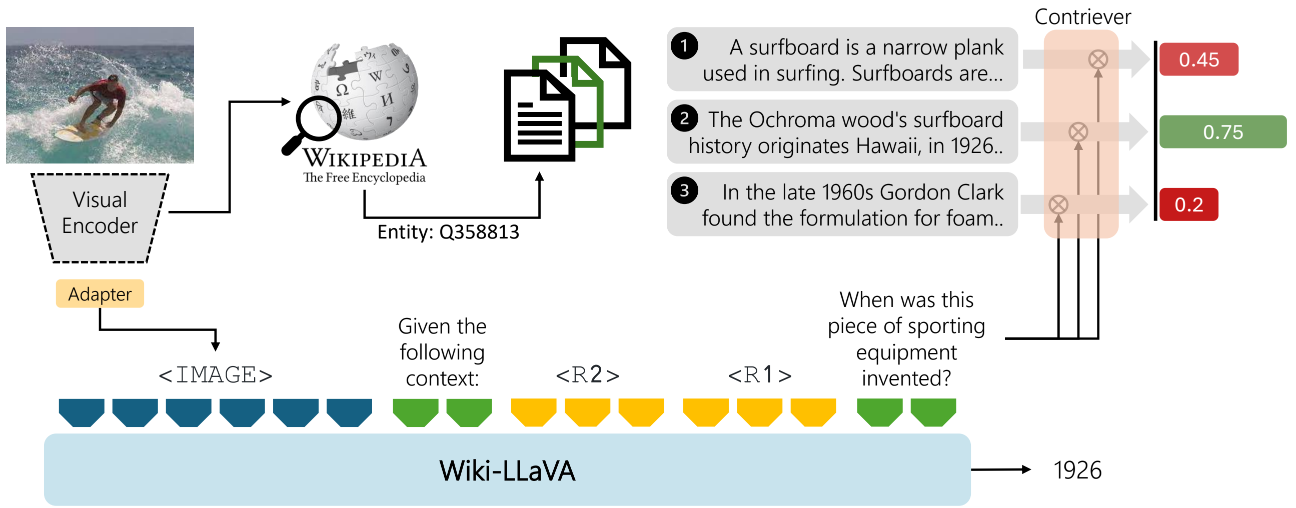

Contriever
1 A surfboard is a narrow plank used in surfing. Surfboards are...
2 The Ochroma wood's surfboard history originates Hawaii, in 1926..
3 In the late 1960s Gordon Clark found the formulation for foam..
0.45
0.75
0.2
WIKIPEDIA
The Free Encyclopedia
Visual Encoder
Adapter
Entity: Q358813
When was this piece of sporting equipment invented?
<IMAGE>
Given the following context:
<R2>
<R1>
Wiki-LLaVA
1926

Figure 2. Overview of the architecture of Wiki-LLaVA, which augments a multimodal LLM with external knowledge through a hierarchical retrieval pipeline.

3. Proposed Method

Our goal is to equip Multimodal LLMs (MLLMs) with the ability to answer complex and specific questions that cannot be addressed solely through the image content and pre-trained knowledge. To achieve this, we propose Wiki-LLaVA, which integrates external knowledge derived from an external memory into the LLaVA model, without significantly altering its design. Instead, we augment the capabilities of the model by incorporating retrieval information as additional input context. Overall, Wiki-LLaVA comprises three components, as shown in Fig. 2: a visual encoder, which is employed to provide the MLLM with visual context and as a query to retrieve from an external knowledge base, the knowledge base itself (*e.g.*, Wikipedia), and a hierarchical retrieval module which retrieves relevant documents and passages from the external knowledge base, to be employed as additional context for the MLLM.

3.1. Knowledge-based Augmentation

Multimodal integration and autoregressive generation. An MLLM usually takes as input a multimodal input query, comprising both image and text, and generates a textual output in an autoregressive manner. Formally, the architecture is trained to model a probability distribution $p(w_t|I, w_0, w_1, ..., w_{t-1}, \theta)$, where θ denotes the parameters of the model, I represents an input image, and $w_0, .., w_{t-1}$ denotes the textual prompt. The textual prompt usually includes a pre-defined system-level prompt and a question related to the input image, given by the user. Clearly, a standard MLLM can only rely on the user prompt, the input image, and the knowledge stored in its internal parameters (*i.e.*, θ) to accommodate requests, thus limiting its ability to answer questions that rely on external knowledge.

In the rest of the paper, we employ LLaVA [24] as our reference MLLM. LLaVA exploits the capabilities of a pre-trained LLM (*i.e.*, Vicuna [6]) and a pre-trained visual model (*i.e.*, a CLIP-based visual encoder [33]), which are interconnected through an MLP adapter, in charge of converting CLIP features to dense input tokens. For an input image I, therefore, LLaVA utilizes a pre-trained CLIP visual encoder E_v, extracts a dense grid of visual features $Z_v = E_v(I)$, which is then projected via a learnable MLP to produce a sequence of dense embedding tokens $v_o, v_1, ..., v_N$. Finally, these are prepended to the system prompt, and the full sequence of visual and textual tokens is then given as input to the LLM component of the model.

Augmentation with external knowledge. To augment the MLLM with external knowledge, we enrich the input context by injecting relevant textual data from an external memory composed of documents. Formally, the distribution of the MLLM is conditioned on additional textual retrieval-knowledge tokens, leading to

$$p(w_t|\overbrace{v_o, v_1, ..., v_N,}^{\text{Visual tokens}} \underbrace{w_0, w_1, ..., w_{t-1},}_{\text{System + user prompt}} \overbrace{e_0, e_1, ..., e_\tau}^{\text{External memory tokens}}), \tag{1}$$

where $e_0, ..., e_\tau$ represents the added tokens retrieved from the external memory. Differently from the standard formulation of MLLMs, by enriching the input context we allow the model to generate more specific answers by exploiting tokens retrieved from the memory.

Hierarchical retrieval from an external memory. The external memory comprises a collection of (document, image, text-title) triplets taken from documents, denoted as $\mathcal{D} = \{(d_i, t_i)_i\}$. Within this memory, we conduct a hierarchical two-step search to retrieve appropriate information.

Initially, we locate the most pertinent document, followed by identifying the relevant passage inside a particular document, which is subsequently exploited as additional input context in the MLLM.

In the first stage, given an input query image I we perform an approximate k-nearest neighbor search into the external memory, using document titles as retrievable keys. The similarity between the query image and the text titles is modeled as the inner product between their respective embeddings, which are computed through the visual and textual CLIP encoders (*i.e.*, E_v and E_t), as follows:

$$\text{sim}(I_i, t_i) = E_v(I) \cdot E_t(t_i)^T. \qquad (2)$$

Then, the knowledge retriever returns the top-k documents associated with the most relevant items retrieved using the aforementioned procedure.

Retrieving document passages. In the second step, we analyze each of the retrieved documents to identify the most relevant passages corresponding to the user's question. Each document is defined as a sequence of chunks, denoted as $d_i = [c_{i_0}, .., c_{i_T}]$, and, given the input question, we retrieve the chunks with the highest similarity to the question. We employ the Contriever architecture [15] to embed each chunk of the selected document, along with the query (*i.e.*, the question provided by the user), and compute the similarity as an inner product between embeddings. By retrieving the n most appropriate passages inside each of the retrieved documents, overall we obtain $k \cdot n$ passages.

Context enrichment. Once we find the most relevant chunks, we employ their raw contents as an additional input to the MLLM. Specifically, the final prompt that we employ includes the image tokens, the retrieved raw chunks, the system-level prompt, and the user question. Formally, considering three retrieved passages, the final prompt is defined as follows:

```
<IMAGE>\nGiven the following context:\n
       <R1>\n<R2>\<R3>\n <QUESTION>
    Give a short answer.  ASSISTANT:
```
(3)

3.2. Training

While the aforementioned approach could work in a zero-shot fashion, using the original weights θ of the pre-trained MLLM, we also investigate the case of fine-tuning the model to augment its capabilities of exploiting retrieved passages. In particular, in this case, the model is trained on pairs of questions and ground-truth answers requiring external knowledge. As this would potentially reduce the capabilities of the MLLM on tasks not requiring external knowledge (*i.e.*, all the other tasks on which the model has been originally trained), we apply a data mixing approach in which ground-truth pairs requiring external knowledge are mixed with ground-truth pairs not requiring external knowledge in the same mini-batch.

4. Experiments

In this section, we first introduce the experimental settings, describing the datasets employed, the evaluation protocol, and the implementation and training details used to perform the experiments. Then, we present our experimental results, analyzing the effectiveness of CLIP fine-tuning and evaluating how it is possible to incorporate retrieved knowledge in an MLLM. Finally, limitations of the proposed approach and possible future works are reported.

4.1. Datasets

Encyclopedic-VQA [28]. The dataset contains around 221k question-answer pairs associated with 16.7k different fine-grained entities, with up to 5 images representing the same entity. Overall, there are more than 1M triplets composed of an image, a question, and the corresponding answer. Fine-grained entities and related images are extracted from iNaturalist 2021 [43] and Google Landmarks Dataset V2 [45], which are associated with the corresponding Wikipedia article. Questions are divided into four different categories, namely single-hop, automatically generated, multi-answer, and two-hop. In particular, single-hop questions have been manually annotated and a single Wikipedia article is needed to answer them. Automatically generated questions are similar to the single-hop questions but have been generated by automatic models. Multi-answer questions, instead, can be answered with a list of terms, but always refer to a single fine-grained entity. Finally, two-hop questions require two retrieval steps to answer them. The dataset also comes with a knowledge base composed of 2M Wikipedia articles, suitable for answering dataset questions.

Dataset triplets are divided into training, validation, and test splits respectively composed of 1M, 13.6k, and 5.8k samples. In our experiments, we employ the training split to fine-tune the LLaVA model and report the results on the test set of the dataset. During testing, we filter out two-hop questions resulting in 4,750 test triplets.

InfoSeek [5]. The dataset contains 1.3M image-question-answer triplets corresponding to around 11k different entities (*i.e.*, Wikipedia articles). The vast majority of questions have been obtained with an almost entirely automatic procedure, by filling human-authored templates with knowledge triples from Wikidata. In this case, images are derived from the OVEN dataset [12]. Triplets are divided into training, validation, and test sets, with around 934k, 73k, and 348k samples respectively. At the time of the submission, the ground-truth answers and entities from the test set were not available. Therefore, we report our results on the validation split. Both validation and test sets contain questions related to new entities not included in the training split and questions not seen during training.

Along with image-question-answer triplets, a knowledge base composed of 6M Wikipedia entities is provided. In our experiments, we consider a randomly extracted subset of 100k entities, in which we guarantee the presence of the 6,741 entities associated with questions from the training and validation splits.

4.2. Implementation Details

LLaVA fine-tuning. We employ two distinct fine-tuning approaches, with each being exclusively applied to one of the datasets. In order to maintain the performance of the LLaVA model on well-established MLLM datasets, we supplement fine-tuning data with samples from the LLaVA-Instruct dataset [24]. Specifically, given its size of 158k, we double the probability of having examples from this dataset in each mini-batch. To reduce the number of trainable parameters, we train using low-rank adapters [11] with a total batch size of 512 samples.

Retrieval. Textual documents sourced from Wikipedia content are embedded using the Contriever architecture [15], segmenting the text into chunks of 600 characters each. Furthermore, for streamlined efficiency, the process involves utilizing a single visual encoder. Specifically, following the LLaVA architecture [24], we employ the CLIP ViT-L/14@336 backbone to embed images to give as input to the MLLM, while simultaneously leveraging it to extract query visual features in the initial hierarchical retrieval step, facilitating the integration of an external memory component.

To perform entity retrieval, we employ approximate *k*NN search rather than exact *k*NN search because it significantly improves the computational speed of the entire pipeline. To this aim, we employ the Faiss library [18] and a graph-based HNSW index with 32 links per vertex.

4.3. Evaluation Protocol

We evaluate our models in two settings: without external knowledge base and with external knowledge base. The former means that we ask the model to directly answer a visual question, by solely relying on the competencies learned during pre-training and/or fine-tuning. On the other hand, in the latter setting, we leverage the proposed hierarchical retrieval method to search for additional information in the external knowledge base. In practice, this is represented by two dumps of Wikipedia comprehending 2M and 100k pages, respectively for Encyclopedic-VQA and InfoSeek. Concerning the evaluation metrics, we report the accuracy over the Encyclopedic-VQA test split and the InfoSeek validation split, following the official evaluation scripts provided along with the datasets.

Dataset	**KB**	R@1	R@10	R@20	R@50
Encyclopedic-VQA	2M	3.3	9.9	13.2	17.5
InfoSeek	100k	36.9	66.1	71.9	78.4

Table 1. Entity retrieval results on the Encyclopedic-VQA test set and InfoSeek validation set. To comply with the visual encoder employed in LLaVA, all results are obtained using CLIP ViT-L/14@336.

4.4. Experimental Results

Analyzing CLIP performance. We start by evaluating entity retrieval results using CLIP. In this setting, we consider images from the Encyclopedic-VQA test set and InfoSeek validation set and measure the CLIP ability to find the correct entity within the knowledge base of each respective dataset (*i.e.*, composed of 2M entries for Encyclopedic-VQA and 100k entries for InfoSeek). As previously mentioned, we perform retrieval using images as queries and Wikipedia titles as retrievable items.

Results are reported in Table 1 in terms of recall@k (R@k) with $k = 1, 10, 20, 50$ which measures the percentage of times the correct entity is found in the top-k retrieved elements. Notably, correctly retrieving the Wikipedia entity associated with the input image strongly depends on the size of the employed knowledge base. In fact, when using 100k items, as in the case of InfoSeek, the correct entity is retrieved as the first item 36.9% of the time and among the top-10 66.1% of the time. Instead, when using a significantly larger knowledge base as in the case of Encyclopedic-VQA, which contains 2M items, retrieval results are significantly lower with 3.3% and 9.9% respectively in terms of R@1 and R@10.

Results on Encyclopedic-VQA and InfoSeek. We then report visual question-answering results in Table 2. We include the performance of zero-shot models like BLIP-2 [21], InstructBLIP [8], and the LLaVA-1.5 baseline model [24], which are not fine-tuned on the considered datasets and that do not leverage the external knowledge base. Moreover, we consider the accuracy results of LLaVA-1.5 when fine-tuned on the training set of Encyclopedic-VQA and InfoSeek, but not augmented with retrieved context. The results of our approach (*i.e.*, Wiki-LLaVA) are reported both in the standard setting in which CLIP is used to retrieve the most representative entity from the knowledge base and in its *oracle* version, which employs the entity corresponding to the input image-question pair. For both cases, we consider a different number n of retrieved textual chunks, all corresponding to the top-1 (or ground-truth) entity. When employing CLIP, we also vary the number k of retrieved entities (*i.e.*, $k = 1, 2, 3$) using $n = 1$ when k is greater than 1. This choice is given by the maximum context length that Vicuna takes as input, which is set to 2,048 tokens.

					Enc-VQA		InfoSeek		
Model	**LLM**	**KB**	k	n	Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot Models									
BLIP-2 [21]	Flan-T5$_{XL}$	✗	-	-	12.6	12.4	12.7	12.3	12.5
InstructBLIP [8]	Flan-T5$_{XL}$	✗	-	-	11.9	12.0	8.9	7.4	8.1
LLaVA-1.5 [23]	Vicuna-7B	✗	-	-	16.3	16.9	9.6	9.4	9.5
Fine-tuned Models									
LLaVA-1.5 [23]	Vicuna-7B	✗	-	-	23.3	28.5	19.4	16.7	17.9
Wiki-LLaVA	Vicuna-7B	✓	1	1	21.8	26.4	26.6	24.6	25.5
Wiki-LLaVA	Vicuna-7B	✓	1	2	19.9	23.2	29.1	26.3	27.6
Wiki-LLaVA	Vicuna-7B	✓	1	3	17.7	20.3	30.1	27.8	28.9
Wiki-LLaVA	Vicuna-7B	✓	2	1	21.3	25.4	27.8	24.6	26.1
Wiki-LLaVA	Vicuna-7B	✓	3	1	20.5	24.3	27.4	24.5	25.3
Wiki-LLaVA	Vicuna-7B	✓	1	1	34.7	37.2	41.1	41.1	41.1
Wiki-LLaVA	Vicuna-7B	✓	1	2	39.2	40.2	49.1	46.5	47.8
Wiki-LLaVA	Vicuna-7B	✓	1	3	38.5	38.6	52.7	50.3	51.5

Table 2. Accuracy results on the Encyclopedic-VQA test set and InfoSeek validation set. **Yellow color** indicates models employing the CLIP model to perform entity retrieval, while **gray color** indicates the use of ground-truth entities (*i.e.*, oracle). k denotes the number of retrieved entities, and n represents the number of textual chunks retrieved for each entity that are given to the MLLM as additional context.

As it can be seen, zero-shot MLLMs face difficulties in correctly answering the given questions as these models can only rely on the knowledge embedded inside the LLM. When instead using an external knowledge base, the accuracy results significantly increase especially on the InfoSeek dataset with 100k retrievable items. The limited performance of the CLIP model in retrieving the correct entity on larger knowledge bases, instead, leads to a slight degradation of accuracy scores. This is due to the noisy textual passages that are provided to the MLLM as additional external context which, being related to a different entity, often do not contain informative content.

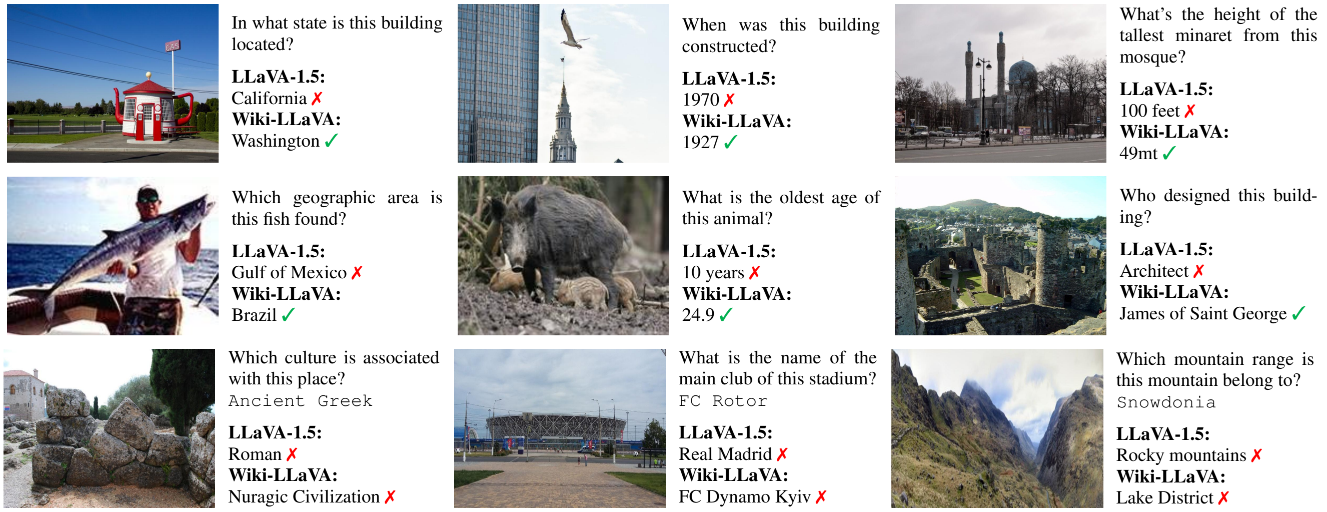
Overall, retrieving passages from different entities does not always help increase the results. Instead, using more than one textual chunk as additional context for the MLLM generally improves the final accuracy on the InfoSeek validation set with an overall improvement of 2.1 and 3.4 accuracy points with $n = 2$ and $n = 3$ respectively. Furthermore, it is worth noting that employing oracle entities significantly boosts the final accuracy. In particular, oracle entities lead to an improvement of 13.8% on Encyclopedic-VQA and 22.6% on InfoSeek, comparing the best-performing configuration with CLIP-based entity retrieval (*i.e.*, $k = 1$ and $n = 1$ for Encyclopedic-VQA and $k = 1$ and $n = 3$ for InfoSeek) with the best performing oracle-based version (*i.e.*, $k = 1$ and $n = 2$ for Encyclopedic-VQA and $k = 1$ and $n = 3$ for InfoSeek). These results confirm the effectiveness of directly employing retrieved passages to augment a pre-trained MLLM and further highlight the importance of having a good entity retrieval model to limit the possibility of feeding the MLLM with irrelevant content.

	Enc-VQA		InfoSeek		
Fine-tuning	Single-Hop	All	Unseen-Q	Unseen-E	All
✗	16.3	16.9	9.6	9.4	9.5
✓	23.4	29.0	17.1	15.0	16.0
✓ + LLaVA-Instruct	23.3	28.5	19.4	16.7	17.9

Table 3. Performance analysis when using the LLaVA-Instruct dataset during fine-tuning. All results are obtained without external knowledge retrieval.

Some qualitative results on sample image-question pairs from Encyclopedic-VQA (first row) and InfoSeek (second row) are reported in Fig. 3, comparing the answers given by Wiki-LLaVA with those coming from the original LLaVA-1.5 model. For completeness, we also report some failure cases (third row) in which both models are not able to correctly answer the given question.

Evaluating the importance of the fine-tuning datasets. As described in Sec. 3.2 and Sec. 4.2, the MLLM fine-tuning is done with a mixture of data containing image-question-answer triples from the Encyclopedic-VQA or InfoSeek training set and visual instruction tuning data from LLaVA-Instruct [24], which has been used to originally fine-tune the LLaVA model. In Table 3, we evaluate the effect of mixing fine-tuning data for the knowledge-based VQA task. In this setting, we only report the results of the fine-tuned models without external knowledge retrieval. Notably, using visual instruction tuning data can help to regularize the fine-tuning phase on the InfoSeek dataset, leading to an overall improvement of 1.9 accuracy points compared to the model fine-tuned only on image-question-answer triplets from the training set of the dataset. On



In what state is this building located?
LLaVA-1.5: California ✗
Wiki-LLaVA: Washington ✓

When was this building constructed?
LLaVA-1.5: 1970 ✗
Wiki-LLaVA: 1927 ✓

What's the height of the tallest minaret from this mosque?
LLaVA-1.5: 100 feet ✗
Wiki-LLaVA: 49mt ✓

Which geographic area is this fish found?
LLaVA-1.5: Gulf of Mexico ✗
Wiki-LLaVA: Brazil ✓

What is the oldest age of this animal?
LLaVA-1.5: 10 years ✗
Wiki-LLaVA: 24.9 ✓

Who designed this building?
LLaVA-1.5: Architect ✗
Wiki-LLaVA: James of Saint George ✓

Which culture is associated with this place?
`Ancient Greek`
LLaVA-1.5: Roman ✗
Wiki-LLaVA: Nuragic Civilization ✗

What is the name of the main club of this stadium?
`FC Rotor`
LLaVA-1.5: Real Madrid ✗
Wiki-LLaVA: FC Dynamo Kyiv ✗

Which mountain range is this mountain belong to?
`Snowdonia`
LLaVA-1.5: Rocky mountains ✗
Wiki-LLaVA: Lake District ✗

Figure 3. Qualitative results on sample image-question pairs from Encyclopedic-VQA (first row) and InfoSeek (second row) comparing the proposed approach with the original LLaVA-1.5 model. Some failure cases are shown in the third row with the corresponding ground-truth.

	MME		**MMMU**	**MMB**	**POPE**	
Fine-tuning	Cogn	Perc	Acc	Acc	Acc	F1
-	355.7	1513.3	35.1	71.6	86.9	85.8
Enc-VQA	200.7	802.8	36.6	67.7	72.9	63.4
Enc-VQA + LLaVA-Instruct	290.0	1170.1	36.6	70.4	87.2	86.6
InfoSeek	296.8	1377.2	35.2	71.7	82.0	79.6
InfoSeek + LLaVA-Instruct	341.3	1438.9	35.6	71.1	85.8	84.2

Table 4. Performance preservation analysis with respect to the original LLaVA-1.5 model (first row) on diverse benchmarks for MLLM evaluation.

Encyclopedic-VQA, instead, training with instruction tuning data does not lead to performance improvement although without degrading the original results.

Preservation of LLaVA performance. Finally, we analyze the impact of LLaVA fine-tuning on knowledge-based VQA datasets when evaluating the model on common MLLM evaluation benchmarks [3]. In particular, we include results on MME [9] which contains image-question pairs covering 14 different tasks grouped in two macro-categories (*i.e.*, cognition and perception), MMMU [47] that is composed of multiple-choice and open-ended questions possibly interleaved with one or more images and extracted from diverse university textbooks and online courses, MMBench (MMB) [26] that includes multiple-choice questions across 20 different domains, and POPE [22] that is focused on evaluating object hallucinations and comprises binary classification entries, each related to an image. More details about the evaluation metrics and number of samples can be found in the original paper of each dataset.

Results are shown in Table 4 comparing the original LLaVA model with the two fine-tuned versions on Encyclopedic-VQA and InfoSeek, with and without the use of visual instruction tuning data. Overall, employing samples from the LLaVA-Instruct dataset can better preserve the results of the original model, only partially degrading the performance on the considered benchmarks compared to the original model. While the most significant deterioration is achieved on the MME dataset, in the other settings the original performances are better preserved, also leading to a slight improvement on MMMU and POPE benchmarks compared to the LLaVA-1.5 results.

4.5. Limitations and Future Works

While our work provides an initial step towards MLLM which can properly exploit external multimodal data, it is worthwhile mentioning that significant research is needed in two directions. The fist is defining proper embedding spaces in which documents can be retrieved from questions and input images, so as to improve the performance of the higher level of our hierarchical retrieval. The second is modeling an efficient and sustainable paradigm to select from one or more documents. Here, the challenge is to increase the capability of the MLLM of distinguishing the appropriateness of retrieved items. This point might also require novel architectural design, which might go beyond the pure inclusion of retrieved items in the context. Regardless of its current limitations, our research testifies the potential of adding multimodal external knowledge to a MLLM and inherits all the advantages of retrieval-augmented approaches, such as the adaptability to different domains and the loosely-coupled relationship between pre-trained information and retrievable data.

5. Conclusion

We have presented Wiki-LLaVA, an architecture for augmenting an existing MLLM with external knowledge. Our proposal leverages an external knowledge source of documents to improve the effectiveness of an MLLM when tasked with questions and dialogues. In particular, we devise a hierarchical architecture for retrieving documents and eliciting selected parts to be included in the MLLM input context. Extensive experiments demonstrate the effectiveness of the proposed solution, and its capability to maintain the proficiency of the MLLM across different tasks.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work has been conducted under two research grants, one co-funded by Leonardo S.p.A. and the other co-funded by Altilia s.r.l., and supported by the PNRRM4C2 project "FAIR - Future Artificial Intelligence Research", funded by the European Commission, and by the PNRR project "Italian Strengthening of Esfri RI Resilience" (ITSERR) funded by the European Union - NextGenerationEU (CUP B53C22001770006).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*, 2022. 1, 2

[2] Manuele Barraco, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning. In *ICCV*, 2023. 2

[3] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The (R)Evolution of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2402.12451*, 2024. 2, 7

[4] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning. In *CVPR*, 2022. 2

[5] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *EMNLP*, 2023. 2, 4

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023. 1, 2, 3

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*, 2022. 1, 2

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*, 2023. 2, 5, 6

[9] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2023. 7

[10] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval Augmented Language Model Pre-Training. In *ICML*, 2020. 2

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[12] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. In *ICCV*, 2023. 4

[13] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. REVEAL: Retrieval-Augmented Visual-Language Pre-Training With Multi-Source Multimodal Knowledge Memory. In *CVPR*, 2023. 2

[14] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language Is Not All You Need: Aligning Perception with Language Models. *arXiv preprint arXiv:2302.14045*, 2023. 2

[15] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv preprint arXiv:2112.09118*, 2021. 2, 4, 5

[16] Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. Select, Substitute, Search: A New Benchmark for Knowledge-Augmented Visual Question Answering. In *SIGIR*, 2021. 2

[17] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active Retrieval Augmented Generation. *arXiv preprint arXiv:2305.06983*, 2023. 2

[18] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale Similarity Search with GPUs. *IEEE Trans. on Big Data*, 7(3):535–547, 2019. 5

[19] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding Language Models to Images for Multimodal Inputs and Outputs. In *ICML*, 2023. 2

[20] Paul Lerner, Olivier Ferret, and Camille Guinaudeau. Cross-modal Retrieval for Knowledge-based Visual Question Answering. *arXiv preprint arXiv:2401.05736*, 2024. 2

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 5, 6

[22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2305.10355*, 2023. 7

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 2, 6

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023. 2, 3, 5, 6

[25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 2

[26] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*, 2023. 7

[27] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*, 2019. 2

[28] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic VQA: Visual Questions About Detailed Properties of Fine-Grained Categories. In *ICCV*, 2023. 2, 4

[29] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. 2

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training Language Models to Follow Instructions with Human Feedback. In *NeurIPS*, 2022. 1, 2

[31] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. *arXiv preprint arXiv:2311.16254*, 2024. 2

[32] Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A Crook, Yifan Ethan Xu, Xin Luna Dong, Christos Faloutsos, Lei Li, Babak Damavandi, and Seungwhan Moon. SnapNTell: Enhancing Entity-Centric Visual Question Answering with Retrieval Augmented Multimodal LLM. *arXiv preprint arXiv:2403.04735*, 2024. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 2021. 2, 3

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(140):1–67, 2020. 1

[35] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. SmallCap: Lightweight Image Captioning Prompted With Retrieval Augmentation. In *CVPR*, 2023. 2

[36] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-Augmented Transformer for Image Captioning. In *CBMI*, 2022. 2

[37] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *CVPR*, 2023. 2

[38] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *ECCV*, 2022. 2

[39] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to Summarize with Human Feedback. In *NeurIPS*, 2020. 2

[40] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford Alpaca: An Instruction-Following LLaMA Model, 2023. 2

[41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[42] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal Few-Shot Learning with Frozen Language Models. In *NeurIPS*, 2021. 2

[43] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking Representation Learning for Natural World Image Collections. In *CVPR*, 2021. 4

[44] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and Benchmarking Universal Multimodal Information Retrievers. *arXiv preprint arXiv:2311.17136*, 2023. 2

[45] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *CVPR*, 2020. 4

[46] Zilin Xiao, Ming Gong, Paola Cascante-Bonilla, Xingyao Zhang, Jie Wu, and Vicente Ordonez. Grounding Language Models for Visual Entity Recognition. *arXiv preprint arXiv:2402.18695*, 2024. 2

[47] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502*, 2023. 7

[48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2