Quantifying Memorization and Parametric Response Rates in Retrieval-Augmented Vision-Language Models

Peter Carragher*, **Abhinand Jha**†, **R Raghav**, and **Kathleen M. Carley**

Carnegie Mellon University
Pittsburgh, PA 15217

Abstract

Large Language Models (LLMs) demonstrate remarkable capabilities in question answering (QA), but metrics for assessing their reliance on memorization versus retrieval remain underdeveloped. Moreover, while finetuned models are state-of-the-art on closed-domain tasks, general-purpose models like GPT-4o exhibit strong zero-shot performance. This raises questions about the trade-offs between memorization, generalization, and retrieval. In this work, we analyze the extent to which multimodal retrieval-augmented VLMs memorize training data compared to baseline VLMs. Using the WebQA benchmark, we contrast finetuned models with baseline VLMs on multihop retrieval and question answering, examining the impact of finetuning on data memorization. To quantify memorization in end-to-end retrieval and QA systems, we propose several proxy metrics by investigating instances where QA succeeds despite retrieval failing. In line with existing work, we find that finetuned models rely more heavily on memorization than retrieval-augmented VLMs, and achieve higher accuracy as a result (72% vs 52% on WebQA test set). Finally, we present the first empirical comparison of the parametric effect between text and visual modalities. Here, we find that image-based questions have parametric response rates that are consistently 15-25% higher than for text-based questions in the WebQA dataset. As such, our measures pose a challenge for future work, both to account for differences in model memorization across different modalities and more generally to reconcile memorization and generalization in joint Retrieval-QA tasks.

1 Introduction

The increasing reliance on LLMs for multimodal tasks across far-reaching sectors such as healthcare, finance, and manufacturing underscores the need to assess the accuracy and reliability of the information they generate. Vision-Language Models (VLM) have achieved state-of-the-art (SoTA) performance on Visual Question-Answering (VQA) benchmarks, and these models often utilize Retrieval-Augmented Generation (RAG) to maintain factual accuracy and relevance in a dynamic information environment. However, this has led to uncertainty in the information the LLM bases its answer on in situations where it may choose between parametric memory and retrieved sources. When models rely on memorized information instead of dynamically retrieving information, they may inadvertently propagate outdated or incorrect information, causing serious legal and ethical risks and undermining trust and reliability in AI systems (Huang et al., 2023).

Despite these concerns, the way that Vision-Language models (VLMs) memorize and retrieve information, particularly in complex multimodal tasks, remains under-explored. Instead, survey studies on parametric knowledge conflicts have found that existing research is focused on reasoning capabilities of unimodal large language models (LLMs) and retrieval augmented generation systems (RAG) (Xu et al., 2024a). Particularly in the context of multimodal retrieval and multihop reasoning, few studies analyze the tradeoff between finetuning for specialized tasks and zero-shot prompting for general-purpose vision-language capabilities. A lack of consensus on how to approach this tradeoff motivates the development of measures to quantify reliance on parametric memory, as well as metrics for quantifying the potential performance impact of extending LLMs with RAG systems.

To address this gap, we investigate how multimodal QA models balance accuracy with memorization on the WebQA benchmark. We compare finetuned multimodal systems against zero-

*Correspondence: petercarragher@cmu.edu

†Work done during affiliation with Carnegie Mellon University, now at Google.

Proceedings of the First Workshop on Large Language Model Memorization (L2M2), pages 127–141

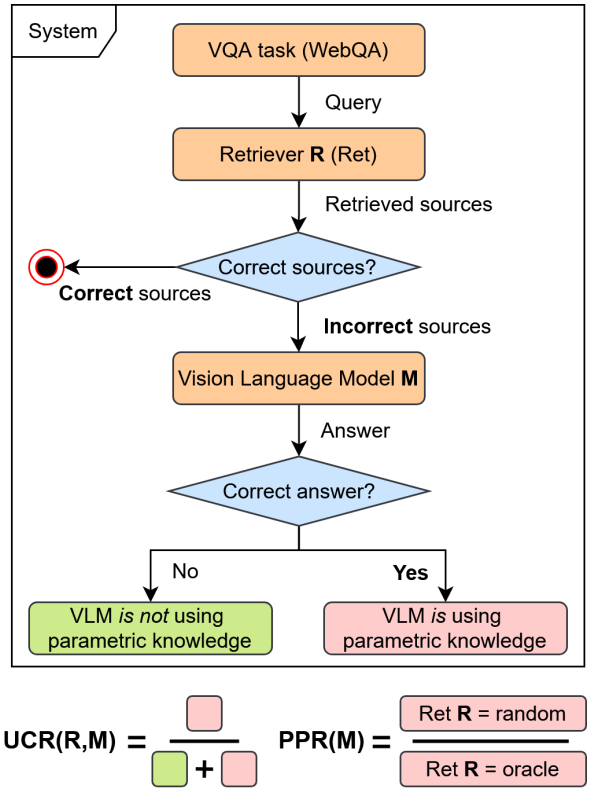


Figure 1: PPR and UCR metrics are derived from the interaction between retrieved sources and QA model.

shot VLMs, analyzing how retrieval performance influences QA accuracy. In particular, we focus on cases where retrieval fails, allowing us to measure reliance on parametric memory through two proposed metrics—the Parametric Proxy Rate (PPR) which quantifies how much model accuracy is influenced by retrieval quality, contrasting performance in best-case versus worst-case retrieval scenarios, and the Unsupported Correctness Rate (UCR) which measures how often correct QA responses are generated when the retriever fails, providing a proxy for memorization. Figure 1 gives an overview of how these measures are derived for a joint retrieval-QA task.

To enable this analysis, we make several methodological contributions. For the finetuned QA models, we investigate Vision-Transformer (ViT) architectures, which allow for multihop reasoning over multiple sources. To investigate the impact of retrieval performance on trained LMs, we propose a variable-input Fusion-in-Decoder (FiD) model (Tanaka et al., 2023; Suhr et al., 2018), building upon the VoLTA architecture (Pramanick et al., 2022). For the zero-shot case, we build upon previous research on In-Context Retrieval (Ram et al., 2023) by demonstrating that LLMs such as GPT-4o are capable of performing the final ranking step of the retrieval process. In doing so, we find that GPT-4o, a general-purpose LLM, achieves SoTA performance on the WebQA task, outperforming existing finetuned RAG models by a significant margin (7% higher accuracy). Crucially, our results reveal that while retrieval-augmented models reduce memorization, the training paradigm plays an important role. Finetuned models exhibit higher reliance on parametric memory, whereas zero-shot RAG approaches have lower memorization scores at the cost of accuracy. While retrievers the improve performance zero-shot VLMs to some degree, as is the case for unimodal systems there is yet no silver bullet in the tradeoff between model generalization and specialization.

Finally, we investigate differences in parametric response rates between text-based and image-based questions from the WebQA dataset. We find that models are capable of answering image-based questions based on parametric knowledge 15-25% more often than they are for text-based sources. In many cases, this means that parametric responses are twice as likely for image-based questions as for text-based ones. Moreover, this result is consistent regardless of the retriever used or whether the QA model was finetuned or not. This finding represents the key empirical contribution from this work—not only is this the first work to measure the parametric effect over image sources, but to the best of our knowledge, it is also the first to present empirical results comparing model memorization tendencies across different modalities. Our findings suggest that the parametric effect may be more pronounced for visual tasks. Future work to validate these findings on additional datasets and problem domains is warranted. We hope that our analysis of model memorization motivates the development of transparent and reliable multimodal AI systems, particularly in applications where the sourcing of up-to-date, factual information from multimodal sources is critical.

2 Related Work

2.1 Multimodal Retrieval Systems

A large body of work on multimodal representations exists (Liu et al., 2022b; Chen et al., 2022b; Radford et al., 2021). CLIP enables the embeddings of text and images into aligned representations by supervised training over image-caption datasets (Radford et al., 2021). More sophisticated

local alignment methods between captions and images using Graph Optimal Transport (GoT) have been proposed (Chen et al., 2020; Petric Maretic et al., 2019). The Universal Vision-Language Dense Retrieval model (UniVL-DR) showed SoTA performance on the WebQA retrieval task (Liu et al., 2022b) by using hard-negative sampling for constrastive learning. In this work, we compare UniVL-DR and CLIP embeddings as competing retrieval systems.

2.2 Multihop Language Models

A wealth of research exists on multimodal Vision-Language tasks and multihop language decoders. (Tanaka et al., 2023) propose a Fusion-in-Decoder (FiD) architecture for multihop reasoning over images. Utilizing advances in local alignment (Chen et al., 2020), VoLTA model combines graph representations of input questions and source images (Pramanick et al., 2022). For compatibility with retriever modules, we extend VoLTA with support for a variable number of input sequences.

More recently, the increasing context windows of VLMs enables them to demonstrate multihop reasoning abilities (Liu et al., 2024; Abdin et al., 2024; Wang et al., 2024). Recent work has found that not only are LLMs capable of determining when they should forgo their parametric memory and use a retriever module (Labruna et al., 2024), they are also capable of "In-context Retrieval" (Ram et al., 2023). Here, retrieved sources are used for grounded text generation by simply prepending the sources into the input prompt. We expand upon this idea, adapting it to a multimodal setting with VLMs, and report our findings.

2.3 The Parametric Effect

There is a wealth of research on reliance on parametric memory for unimodal QA tasks (Galway et al., 2024; Xu et al., 2024b; Longpre et al., 2022; Neeman et al., 2022; Hong et al., 2024; Chen et al., 2022a). Here, the entity replacement framework (Longpre et al., 2022; Neeman et al., 2022) is used to invalidate parametric memory by explicitly crafting knowledge conflicts between input sources and parametric memory (Xu et al., 2024b; Hong et al., 2024; Chen et al., 2022a). As such, these studies guarantee that manipulated input sources no longer entail the expected labels, and focus on evaluating LLMs in isolation without using retrieval systems.

In contrast, we do not make the same guarantees, and our proxy measures are premised upon the key assumption that incorrectly retrieved sources *do not entail* the correct answer. Our focus is on developing proxy metrics for the parametric effect that do not require such involved source manipulation processes. Rather, building upon prior work on unimodal LLMs (Soudani et al., 2024), these metrics compare the performance of finetuned VQA models with RAG systems.

3 WebQA Dataset

The WebQA dataset (Chang et al., 2022) uses a two-step design; retrieval followed by QA. First, given the question Q and all sources S, we retrieve the set of relevant sources, S'. Using these sources we then generate an answer A'. The following is passed to the QA classifier:

$$< [CLS], s'_0, [SEP], \ldots, s'_n, [SEP], Q > \quad (1)$$

We include only those questions that require either one (n = 12,027) or two (n = 9,438) image sources. For a breakdown of question categories and their keywords, see 8 in the appendix. The remaining questions use only text sources (n = 20,267). Our final analysis of unimodal vs multimodal parametric effects uses this portion of the dataset to evaluate memorization on text sources.

As opposed to WebQA, open-domain VQA tasks such as OK-VQA (Marino et al., 2019) and HotpotQA (Yang et al., 2018) do not provide candidate sources S and source labels S^*, and as a result are incompatible with or measures (see section 5). Moreover, while we do evaluate model performance on VQA datasets (NLVR2 (Suhr et al., 2018) and VQAv2 (Goyal et al., 2017)), these tasks lack a retrieval step and so are only useful for QA model selection (see section A.1).

4 Methodology

As WebQA is a joint retrieval and QA task, we develop several QA methods and retrieval methods separately. Using the best-performing QA model, we then evaluate end-to-end retrieval and VQA performance and investigate the factors that affect the parametric effect.

4.1 Question Answering

Vision-Language Model For two-image questions, the WebQA finetuned VLP baseline (Zhou et al., 2020) takes as input the concatenation of both sources encodings with the query;

$$< [CLS], s_1, s_2, [SEP], Q > \quad (2)$$

As such, it is an extension of VQA model trained on single-hop VQA-2 (Yu et al., 2023), which takes as input:

$$< [CLS], s, [SEP], Q > \quad (3)$$

We adopt this formulation for finetuning the Qwen2 VLM, using Low-Rank Adaptation (LoRA) to reduce trainable parameters (Hu et al., 2021). We use the same input formulation to evaluate zero-shot performance on GPT-4o. In addition, we evaluate several baseline models from previous works, namely VLP (Zhou et al., 2020), GIT (Wang et al., 2022), GPT-3.5 (Brown et al., 2020), and BLIP-2 (Li et al., 2022). Details of these models are presented in appendix subsection A.6.

Multihop Formulation We hypothesize that multihop tasks, such as WebQA, would benefit from a two-stage reasoning process. The first stage enables multimodal fusion between each input source and the question, and the second stage enables multihop fusion between the embedded multimodal representation of each source, conditioned on the question. Inspired by FiD architectures (Yu et al., 2022), this results in the following input construction:

$$\begin{aligned} concat(&< [CLS], s_1, [SEP], Q >, \\ &< [CLS], s_2, [SEP], Q >) \end{aligned} \quad (4)$$

Multihop Classifier We select the VoLTA framework as the skeleton for encoding joint text and image representations (Pramanick et al., 2022). VoLTA uses Swin-Base (Liu et al., 2021) and RoBERTa-Base (Liu et al., 2019) as respective visual and textual encoders and we adopt the same encoder choices. We jointly encode each image source returned by the retriever with the query and concatenate the resulting embeddings together before sending them to the MLP classifier to predict the keyword answer label. To handle variable input sequences during classification, we pad single image sources with blank images so that all inputs sent into the classifiers have two images. We call this model MultiHop-VoLTA (MH-VoLTA).

We finetune the models using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $1e^{-4}$ and a batch size of 32 samples. We use LoRA to reduce trainable parameters (Hu et al., 2021), and set $r = 8$ and $\alpha = 32$ for the text encoder and $r = 16$ and $\alpha = 16$ for the image encoder, updating only the attention weights. MH-VoLTA is trained until convergence (80 epochs, see Figure 5c in the appendix).

4.2 Retrieval Methods

Dense Retrievers We adopt the pretrained UniVL-DR retriever for source retrieval in our finetuned experiments (Liu et al., 2023) and compare it with baseline CLIP (Radford et al., 2021) and WebQA finetuned CLIP (CLIP-DPR, (Liu et al., 2023)) embeddings. Specifically, we embed all text sources, image sources, and queries using UniVL-DR. For each query, we compute cosine similarity between the query and each of the sources, and use the top two ranked image sources and their captions as input to the QA model.

GPT-4o Ranking We utilize GPT-4o to select sources from the set of distractor sources present in dataset using the prompt in the appendix subsection A.4. This is motivated by previous work in In-Context Retrieval Augmented Language Modeling (In-Context RALM) which demonstrated that LLMs are capable of reasoning over sources without finetuning (Ram et al., 2023).

Upper and Lower Bounds In addition, to investigate the impact of the parametric effect on joint retrieval and QA performance, we also compare performance with a best and worst case retriever. The best case is the oracle retriever, using gold sources provided in the validation set, and the worst case is a random naive retriever, which returns random distractor sources (and so is always incorrect).

5 Evaluation Metrics

We propose measures for evaluating the degree of memorization in QA models (Parametric Proxy Rate) and in end-to-end retrieval-QA systems (Unsupported Correctness Rate), as well as a metric for retriever-QA model compatibility (Retriever Potential Attainment).

5.1 Unsupported Correctness Rate

We propose UCR, a metric to measure the parametric effect in the combined retrieval and QA model. It is formulated as a composition of QA accuracy and retrieval recall. Intuitively, it is the fraction of true positive predictions from the QA model for which there is no retrieval support (i.e. the retrieved sources were incorrect).

Retrieval Recall and QA Accuracy The first stage of the joint task is retrieval, where the recall for retriever R is defined as the fraction of retrieved sources (positives) that are correct (true positives)

with respect to task labels;

$$\text{Recall}_R = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (5)$$

Accuracy is the primary correctness metric for question answering in the WebQA task. Accuracy of model M is determined by comparing a restricted bag of words (bow) vector between the expected (E) and generated (G) answers;

$$\text{Acc}_M = \frac{1}{n}\Sigma[\frac{|\text{bow}_E \cap \text{bow}_G|}{|\text{bow}_E|} == 1] \quad (6)$$

The vocabulary of the vectors is restricted to a specific domain based on the question type; questions are labeled based on these domains which can be yes/no, color, shape, or number. Each category has a pre-defined vocabulary list, given in the appendix.

UCR Using QA accuracy and retrieval recall, we construct UCR, a metric for measuring the parametric effect in a combined retrieval-QA model, which calculates the likelihood $P(Q_1|R_0)$ that the QA model M returns a correct answer (Q_1) given that the retrieval model R failed to return the correct sources (R_0):

$$\begin{aligned}\text{UCR(R,M)} &= \frac{\text{Acc}_M == 1 \cap \text{Recall}_R == 0}{\text{Recall}_R == 0} \\ &= P(Q_1|R_0)\end{aligned} \quad (7)$$

5.2 Oracle-Normalized Retrieval Scores

Using min-max scaling, we define two additional metrics to evaluate joint retrieval QA systems, by normalizing using the oracle retriever (upper bound) and random retriever (lower bound):

$$\hat{X} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

Retriever Potential Attainment RPA quantifies the potential that a retriever has realized when used in a given end-to-end retrieval QA system. The upper bound (1) is given by same QA system's accuracy with oracle sources ($\text{Acc}_M(\text{oracle})$). The lower bound (0) is given by the random negative source retriever ($\text{Acc}_M(\text{random})$), which always retrieves incorrect sources. Note that $\text{Acc}_M(\text{random})$ is similar to the Free Success Rate metric (Lin and Byrne, 2022), which represents the QA model's accuracy with no retrieved sources. We apply random-oracle scaling, where $Acc_M(R)$ denotes the accuracy of QA model M, given sources from retriever R:

$$\text{RPA(R,M)} = \frac{\text{Acc}_M(R) - \text{Acc}_M(\text{random})}{\text{Acc}_M(\text{oracle}) - \text{Acc}_M(\text{random})} \quad (9)$$

Parametric Proxy Rate We postulate that the rate at which a model's performance increases when used in conjunction with increasingly accurate retrievers implies that it is using the retrieved sources effectively, instead of relying on parametric memory. To that end, we present PPR, an additional max scaling measure based off just the upper (oracle) and lower bound (randomized *negatives*) retrievers, which is simply their performance ratio with respect to QA model M:

$$\text{PPR(M)} = \frac{\text{Acc}_M(\text{random})}{\text{Acc}_M(\text{oracle})} \quad (10)$$

6 Results

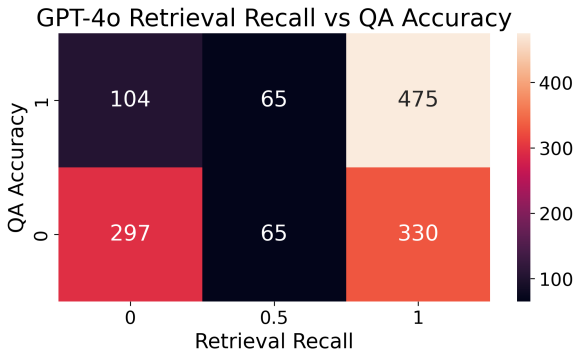
First, we experiment with multiple QA models to determine which generalized LLMs and specialized, finetuned models should be selected for the joint retrieval and QA task. Then, we analyze how performance on the retrieval task impacts QA accuracy and experiment with different combinations of retrieval systems and chosen QA models for the joint task. Finally, we investigate the effect of finetuning on model memorization and compare parametric response rates over WebQA image-based and text-based questions.

6.1 Model Selection

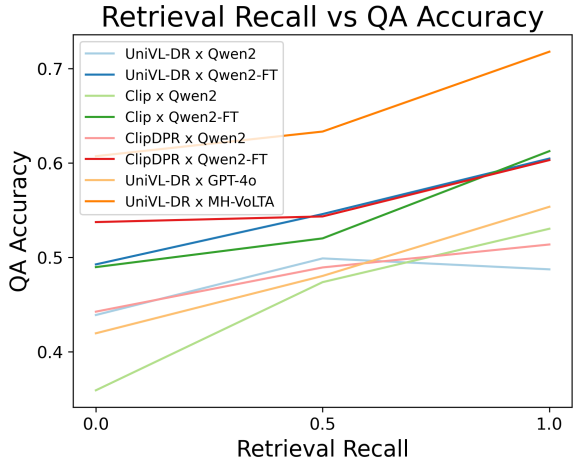
We find that the MH-VoLTA model outperforms all baseline and zero-shot models on the WebQA validation set image questions, including BLIP-2, GIT, VLP, GPT-4o, and GPT-3.5. We also find that MH-VoLTA performance is comparable to VoLTA on the (fixed input) VQA and NLVR2 tasks (section A.1 in the appendix). For a breakdown of model performance by question category on the WebQA dataset, see section A.3 in the appendix.

6.2 Impact of Retrieval on QA

To understand how retrieval and QA systems interact, we investigate the reliance of the QA task on retrieval correctness (Figure 2). We find that when GPT-4o correctly retrieves the relevant sources through In-Context Retrieval, it has a 59% QA accuracy rate. If GPT-4o In-Context Retrieval fails to retrieve the correct sources, the accuracy rate is



(a) Most questions answered correctly by GPT-4o have correctly retrieved sources, as given by low UCR scores: $UCR(\text{GPT}, \text{GPT}) = \frac{104}{104+297} = 0.26$



(b) Retrieving distractor sources decreases QA accuracy.

Figure 2: Across experiments, recall impacts QA.

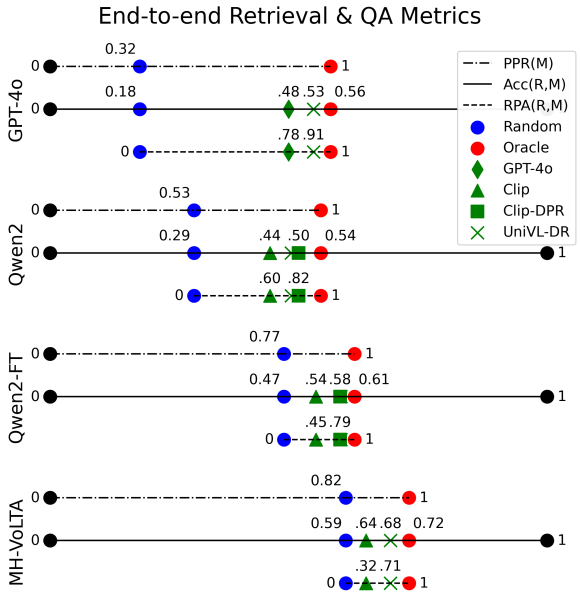


Figure 3: Evaluation metrics on end-to-end retrieval & QA systems: Accuracy (Acc, Equation 6), Parametric Proxy Rate (PPR, Equation 10), and Retriever Attainment (RPA, Equation 9) (denoted by the three lines) for each pairing of retriever R and QA model M. Note that the lines denoting PPR and RPA are rescaled to represent the denominators in the respective equations.

reduced to 26% (Figure 2a). We also find that QA performance drops as the number of retrieved distractors increases and retrieval recall falls, showing that poor retrieval performance adversely affects QA (Figure 2b). This is to say that the QA task is heavily dependent on retrieval performance. However, there do exist correctly answered questions for which incorrect sources are retrieved, and these samples form the basis for the UCR measure of the parametric effect (Figure 2a).

In addition, we contribute error analysis in the appendix that show the differences between In-Context Retrieval using GPT-4o and dense retrieval methods such as UniVL-DR (subsection A.5). In particular, we find that systems that rely on "in-context" retrieval using GPT-4o are limited by query complexity, but approaches that utilize dense retrievers are not. As query complexity increases, In-Context Retrieval degrades QA accuracy (Figure 6), but dense retrievers do not (Figure 7).

6.3 End-to-End Retrieval and QA

The PPR and RPA measures enable a quick comparison of joint retrieval and QA systems, where Figure 3 reveals some interesting trends. We find that of all QA models tested, GPT-4o benefits the most from the use of retrievers—Retriever Potential Attainment (RPA) scores are highest for GPT-4o—while finetuned QA models such as Qwen2-FT and MH-VoLTA receive a lower performance increase as the coupled retrieval system is improved.

GPT-4o also has the best PPR score. That is, GPT-4o has the biggest gap in performance when comparing the worst case (random negative) and best case (source oracle) retrievers, with a PPR of 0.32. In comparison, Qwen2 has higher performance under the random retriever, and as such displays a greater reliance on parametric memory.

There is also a clear trend between finetuning, QA accuracy, and Parametric Proxy Rate (PPR). While finetuned Qwen2 (Qwen2-FT) has improved accuracy vs Qwen2, it's performance on the worst case retriever is surprisingly high (PPR =0.77). This is even more extreme for MH-VoLTA, which obtains both the highest QA accuracy (0.72) and the highest PPR (0.82). The same trend is apparent when evaluating on the WebQA test set, where fine-

Retriever R	Model M	$\text{Acc}_M^{\text{test}}(R)$
UniVL-DR	Qwen2	0.52
UniVL-DR	Qwen2-FT	0.70
Uni-VLDR	GPT-4o	0.73
GPT-4o	GPT-4o	**0.77**

Table 1: QA Accuracy on WebQA test set.

Retriever R	Recall	Unsupported Correctness Rate			
		Qwen	Q-FT	MHV	GPT4
Rand	0.00	0.260	0.449	0.595	0.174
Clip	0.46	0.328	0.467	0.617	–
DPR	0.77	0.420	0.517	0.643	–
UVL	0.80	0.438	0.521	0.616	0.420
GPT	0.65	–	–	–	0.259

Table 2: UCR ($P(Q_1|R_0)$) for each retriever R and QA model M, alongside retrieval recall. DPR denotes Clip-DPR, Q-FT is Qwen2-FT, UVL is UniVL-DR.

tuning Qwen2 improves accuracy (Table 1). Note, PPR cannot be measured on the test set, as labels have not been made public.

6.4 Finetuning and UCR

We find that, while the finetuning process improves accuracy (and in part because of this fact), finetuning exacerbates the parametric effect. Qwen2-FT has a higher Parametric Proxy Rate than the baseline Qwen2 model (PPR, Figure 3), and it has a higher Unsupported Correctness Rate (UCR) than Qwen2 across all retrieval methods tested (Table 2). What's more, the act of finetuning Qwen2 has an outsized effect on UCR when compared with the effect that changing the retriever has. MH-VoLTA represents the extreme case; for each retriever R, $\text{UCR}(R, \text{MH-VoLTA}) > 0.5$, implying that MH-VoLTA is correctly answering the majority of questions for which the retrieval system fails to identify the correct sources.

However, the effect of retrieval on UCR is not negligible, and we find that for a given QA model, UCR increases as retrieval recall increases; i.e. for each model M $\text{UCR}(\text{Rand}, M) < \text{UCR}(\text{Clip}, M) < \text{UCR}(\text{ClipDPR}, M)$ (Table 2). This implies that as the retriever improves, the QA model is more successful on samples that retrieval fails on. This paradox is explained by inaccuracies in the source labels—distractor sources often provide enough context for the QA model to answer correctly. Rather than exposing memorization, this reveals an underlying issue with the source labels in the WebQA dataset, and as such, these measures can be adapted to evaluate the correctness of the joint retrieval-QA benchmarks.

6.5 Multimodal vs Unimodal Memorization

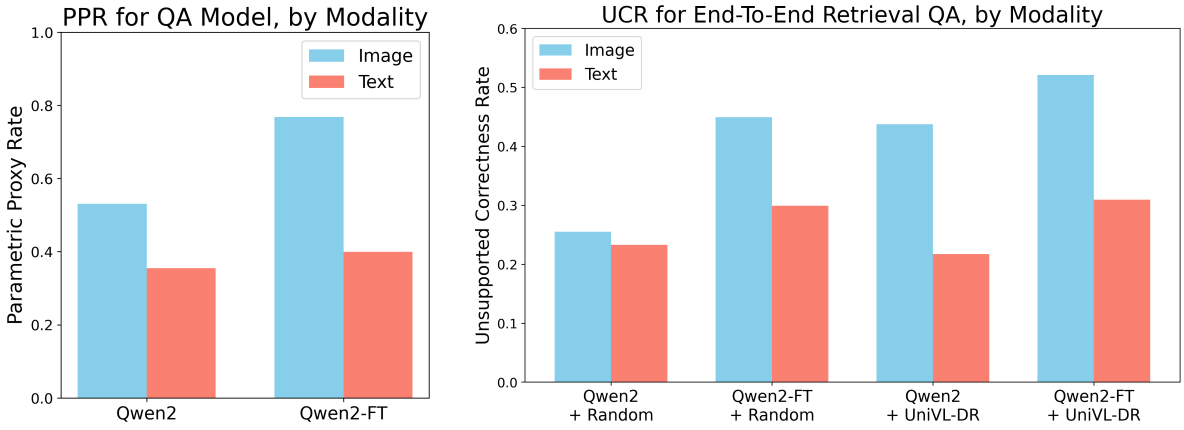
Finally, using the developed metrics and the finetuned and non-finetuned VLM models, we investigate the differences between textual and visual parametric knowledge. Figure 4 reveals that parametric responses are more prevalent for webqa image questions than for webqa text questions. While finetuning results in a minor increase of parametric response rates for the text modality, these rates increase dramatically for the visual modality after finetuning. For example, while UCR on text-based questions increases by $\sim 10\%$ with finetuning, it increases by $\sim 20\%$ for image-based questions. In addition, parametric responses are as much as twice as likely when models are presented with image sources, and these results are consistent across the metrics and models used. For example, for Qwen2-FT, $\text{PPR}_{img} = 0.77$ while $\text{PPR}_{txt} = 0.4$, and when Qwen is paired with the UniVL-DR retriever, the end-to-end system has $\text{UCR}_{img} = 0.44$ and $\text{UCR}_{txt} = 0.22$. These results highlight a modality-based disparity in how memorization manifests in QA models.

7 Discussion

While QA performance is generally predicated upon retrieval success (Figure 2b), there are many cases where retrieval fails and QA succeeds (Figure 2a). These cases form the basis of our quantitative metrics, the Unsupported Correctness Rate (UCR; see Table 2) and Parametric Proxy Rate (PPR; see Figure 3), and with these measures we show that external retrievers significantly reduce the reliance of VLMs on parametric memory. This not only preserves model flexibility but also mitigates the over-specialization common in finetuned systems. However, despite GPT-4o obtaining state-of-the-art performance on the WebQA benchmark using this approach (Table 1)[1], for less powerful VLMs such as Qwen2 the decrease in Parametric Proxy Rate associated with not finetuning the model (PPR: 0.77 -> 0.53) comes at the cost of model accuracy (QA accuracy: 70% -> 52%).

In interpreting these results, it is important to

[1]"Anon_Feb25" @ WebQA leaderboard



(a) PPR(image) > PPR(text)

(b) UCR(image) > UCR(text)

Figure 4: Higher implies greater levels of memorization. Across all metrics, QA models, and retrievers tested, parametric responses are more prevalent for webqa image questions than for webqa text questions.

note that UCR and PPR are proxy measures based upon the key assumption that incorrectly retrieved sources *should* result in incorrect answers from the VQA model. While we can guarantee that distractor sources from the random retriever provide no useful information to the model (i.e. they are randomly sampled negatives), fully addressing this assumption requires modifying the image sources to invalidate the original answer or label. Along these lines, a recent line of research uses constrained image generation to create knowledge conflicts between image sources and parametric memory (Carragher et al., 2025). This builds on research into provoking parametric responses from unimodal LLMs, where the entity replacement methods (Longpre et al., 2022; Neeman et al., 2022) create knowledge conflicts between text sources and parametric memory (Xu et al., 2024b; Hong et al., 2024; Chen et al., 2022a). Entity replacement frameworks for VLMs (Carragher et al., 2025) can make use of object detection (Ravi et al., 2024) and visual attention models (Selvaraju et al., 2020) to allow parametric analysis to move beyond the incorrectness assumption.

Despite this assumption, our measures reveal an interesting interplay between retrieval and parametric responses. By providing insights into end-to-end retrieval and QA systems, UCR can highlight when models are over-reliant on parametric memory. High Retriever Attainment scores (RPA) across Qwen2 and GPT-4o experiments demonstrate that general-purpose VLMs can utilize finetuned retrievers (Figure 3), drawing the need for domain-specific finetuning into question. This work points towards In-Context Retrieval as a particularly promising direction for future research in multimodal systems, if the limitation regarding question complexity can be addressed (A.5).

Crucially, while image manipulation methods are not subject to the incorrectness assumption that the proxy metrics proposed here are, they are not applicable to text sources. Our proxy measures allow for the comparison of how the parametric effect manifests across different modalities in multimodal language models. Herein, we find that parametric responses may be more prominent over image sources as opposed to text sources (Figure 4). Given that the parametric effect is, as of yet, understudied in multimodal setting (Xu et al., 2024a), to the best of the authors knowledge this is the first comparison of parametric effects between modalities. Our finding motivates incorporating the wealth of parametric research on unimodal models into the multimodal domain (Carragher et al., 2025).

Overall, our methodology provides a framework for measuring and mitigating memorization in retrieval-augmented systems, offering new ways to evaluate the quality of retrieval-QA datasets. Future work can apply this framework to improve retrieval-aware finetuning strategies, where LLMs learn when to prioritize retrieved content rather than rely on parametric knowledge (Labruna et al., 2024). Extending our parametric analysis to open-domain multimodal tasks would provide new insights into retrieval dynamics in unrestricted, real-world settings. Finally, by quantifying reliance on parametric memory, researchers can better assess the trade-offs within finetuning and retrieval per

modality, thereby guiding the development of multimodal models that balance generalization with task-specific performance. As retrieval-augmented VLMs continue to scale, our findings highlight the need for multimodal evaluation of parametric responses to ensure safe, effective, and adaptable AI systems.

7.1 Limitations

The UCR analysis presented in Figure 4b is subject to limitations based on the design of the metric. Specifically, if certain question categories have higher retrieval recall, the number of incorrectly retrieved samples for that category will be low. A study of UCR across question categories, including suitable confidence intervals, is warranted.

Our analysis also reveals a paradoxical relationship between retriever recall and UCR that highlights a potential annotation issue within the dataset, prompting a reevaluation of how retrieval-QA benchmarks are constructed (Table 2). While this means that UCR can be used in the evaluation of retrieval tasks to highlight potential false negative annotation issues, as in WebQA (Table 2), annotation inconsistencies in turn affect the reliability of UCR as a sole indicator of retrieval quality. As such, our findings should be validated on additional VQA datasets. To facilitate this, widely used VQA datasets could be augmented with retrieval tasks, such that joint retrieval-QA systems may be evaluated on them.

Finally, given the rapid pace of research in the multimodal space, the WebQA dataset may have already been incorporated into the training data of the VLMs investigated here. While WebQA is not specifically listed among the Qwen2-VL training materials (Wang et al., 2024), many similar datasets are. For models that have been pretrained on the same joint retrieval-QA dataset used to operationalize the measures proposed here, the measures may be more indicative of verbatim memorization, where model output exactly matches the dataset labels. In contrast, our analysis is targeted at the parametric effect, which is a preference for models to reason over parametric knowledge instead of input sources. Disentangling verbatim memorization from the parametric effect is a good avenue for future research, as it represents a more dramatic failure case of model generalization.

8 Conclusion

We demonstrate that retrieval-augmented VLMs have improved performance over general-purpose VLMs, with comparable parametric response rates. However, there is still a substantial performance gap between finetuned and baseline QA models. By introducing UCR and PPR, we provide concrete measures of how incrementally improving retrieval systems mitigates parametric responses. This analysis outlines the interplay between parametric knowledge and external retrieval, highlighting well-known tradeoff between memorization and generalization in the multimodal setting. Finally, we demonstrate that current VLMs have higher parameteric response rates when reasoning over image sources rather than text sources. Our work provides a foundation for future research aimed at refining retrieval mechanisms and ensuring that external sources effectively complement the parametric knowledge of VLMs.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Peter Carragher, Nikitha Rao, Abhinand Jha, R Raghav, and Kathleen M Carley. 2025. Segsub: Evaluating robustness to knowledge conflicts and hallucinations in vision-language models. *arXiv preprint arXiv:2502.14908*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504.

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. 2022a. Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence. *arXiv preprint*. ArXiv:2210.13701 [cs].

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph optimal transport for cross-domain alignment. *Preprint*, arxiv:2006.14744 [cs].

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022b. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Michael Galway, Matteo DiRenzo, Dominik Esposito, Rafael Marchand, and Valentin Grigoriev. 2024. The Mitigation of Excessive Retrieval Augmentation and Knowledge Conflicts in Large Language Models.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Robert Gunning. 1952. The technique of clear writing. *(No Title)*.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why So Gullible? Enhancing the Robustness of Retrieval-Augmented Models against Counterfactual Noise. *arXiv preprint*. ArXiv:2305.01579 [cs].

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Tiziano Labruna, Jon Ander Campos, and Gorka Azkune. 2024. When to retrieve: Teaching llms to utilize information retrieval effectively. *arXiv preprint arXiv:2404.19705*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Preprint*, arXiv:2103.14030.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022a. Universal multi-modality retrieval with one unified embedding space. *arXiv preprint arXiv:2209.00179*.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022b. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations*.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023. Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. *arXiv preprint*. ArXiv:2209.00179 [cs].

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2022. Entity-Based Knowledge Conflicts in Question Answering. *arXiv preprint*. ArXiv:2109.05052 [cs].

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering. *arXiv preprint*. ArXiv:2211.05655 [cs].

Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, and Pascal Frossard. 2019. Got: an optimal transport framework for graph comparison. *Advances in Neural Information Processing Systems*, 32.

Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik Shah, Yann LeCun, and Rama Chellappa. 2022. Volta: Vision-language transformer with weakly-supervised local-feature alignment. *arXiv preprint arXiv:2210.04135*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 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. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint*. ArXiv:2408.00714 [cs].

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359. ArXiv:1610.02391 [cs].

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 12–22.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024a. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge Conflicts for LLMs: A Survey. *arXiv preprint*. ArXiv:2403.08319 [cs].

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. Unified language representation for question answering over text, tables, and images. *arXiv preprint arXiv:2306.16762*.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022. KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. *arXiv preprint*. ArXiv:2110.04330 [cs].

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

A Appendix

A.1 Model Selection Results

We explore baseline methods for the QA task on the WebQA validation set. Table 3 gives results for the baseline models. The MH-VoLTA model outperforms all baseline and zero-shot models on the validation set image questions. However, the extension of the VoLTA model for variable input multi-hop tasks risks a regression in performance on traditional VQA tasks which have fixed-input where the number of input images is constant. To determine MH-VoLTA generalizes from fixed to variable input tasks, we compare performance between two variants of the original VoLTA model, finetuned on one and two image subsets of WebQA, with MH-VoLTA. We find that MH-VoLTA is capable of reasoning over both one and two-image image questions, and it's performance is on-par with VoLTA variants trained on one and two image sources separatelyTable 3. See subsection A.2 for more details on the one and two image VoLTA variants, as well as a breakdown of model performance by question category (Figure 5a). See subsection A.6 for a description of the baseline models used.

Table 3: Model selection results on WebQA validation set (further broken into 1 and 2 image input categories), and the VQAv2 and NLVR2 (NLV) test sets. MH-V denotes MH-VoLTA. See subsection A.6 for model descriptions.

	WebQA Acc			VQA	NLV
Model	All	1 img	2 img	Acc	Acc
MH-VoL	**0.71**	0.72	0.70	73.9	76.5
VoLTA_1	–	**0.77**	–	**74.6**	–
VoLTA_2	–	–	**0.84**	–	**76.7**
GPT-4o	0.56	**0.58**	0.69	–	–
Qwen2	0.54	–	–	–	–
GPT-3.5	0.53	0.41	0.45	–	–
VLP	0.50	0.40	0.42	–	–
GIT	0.42	0.43	0.35	–	–
BLIP-2	0.40	0.37	0.44	–	–

Table 4: MH-VoLTA results and dataset breakdown

	No. of Samples	Accuracy
Single Image	760	0.764
Two Images	576	0.851
Multiple Images	1336	0.799

VQAv2 and NLVR2 As our memorization metrics require that the task be designed in a two-part retrieval and VQA process, this leaves WebQA as the only valid VQA task to evaluate performance on. Here, independently of any external retrieval system, we validate MH-VoLTA performance on two fixed-input VQA datasets (see Table 3). As the other baseline models have been validated in prior works, we do not measure their performance on VQAv2 or NLVR2 again.

We evaluate models VQAv2 (Goyal et al., 2017), a multi-class, single-image VQA dataset, and (Suhr et al., 2018), a binary classification, two-image VQA dataset. These datasets are well-suited to VoLTA classifier architecture. In particular, question categories in VQAv2, along with the associated answer-domains, match well with WebQA, with a substantial portion of both datasets focusing on color, shape, number, and yes/no questions.

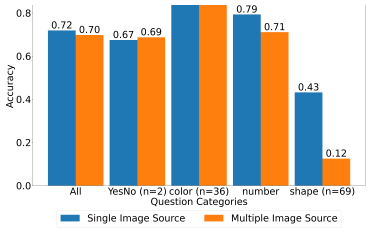
A.2 Multihop VoLTA on one vs two image sources

The results for finetuning VoLTA and MH-VoLTA on the WebQA dataset experiments are provided in Table 4. We explored the application of Multihop-VoLTA in addressing queries based on single images, questions involving two images, and a combination of both single and two-image queries (referred to as multiple images, Figure 5b).

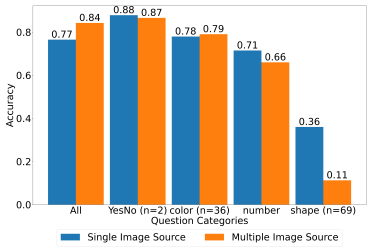
We find that the variable Multihop-VoLTA model (Figure 5a) is en-par with the fixed-input one and two-image VoLTA model variants (Figure 5b). This underscores the stability of our finetuning approach for MH-VoLTA across both training paradigms. The MH-VoLTA models have on the order of 100M parameters, of which 10M are trainable after applying LoRA. All models are trained for 80 epochs on a Nvidia A6000.

A.3 Performance by Question Category

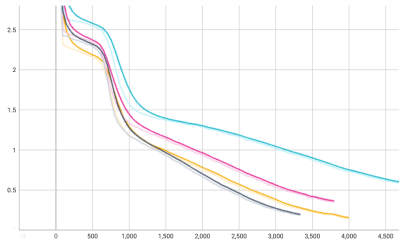
We report the mean accuracy per question category for Multihop-VoLTA in Figure 5a using source retrieval oracles. We find that performance is dependent upon the level of training data available, with the shape category having the least number of samples in the dataset. Question counts per category are as follows; Yes/No (n = 7,320), color (n = 1,830), number (n = 2,118), shape (n = 565). The similarity in results across different question categories reinforces the reliability and stability of our



(a) Performance of the MH-VoLTA classifier by question category and image count.



(b) Performance of fixed input single and two-image VoLTA classifiers.



(c) Convergence of the VoLTA loss function on the WebQA dev set across several MH-VoLTA training runs.

Figure 5: Comparison of the variable MH-VoLTA model (left) vs fixed input VoLTA models (center) across different question categories, ordered by the number of image sources per question. Models converge after 80 epochs (right).

model's performance. For a breakdown of labels per question category, see Figure 8.

A.4 GPT-4o Retrieval Prompt

system: Answer the question in one word. Then list the Fact_ID or Image_ID of all facts used to derive the answer in square brackets.
human: Question: <query>
human: Text Facts: [fact_id_1: fact_1, ..., id_n: fact_n]
human: Image_ID: img_id_1, Caption: img_caption_1
human: [Input_type=image] image_url=url_1
...
human: Image_ID: img_id_m, Caption: img_caption_m
human: [Input_type=image] image_url=url_m

A.5 Robustness Checks: Question Complexity

Complexity Metrics To identify correlations between the complexity of the question, retrieval recall, and QA accuracy, we apply three separate measures to the input questions; Word Count, Flesch-Kincaid Grade Level (Flesch, 2007), and Gunning-Fog Index (Gunning, 1952).

The Flesch-Kincaid Grade Level is a readability metric that evaluates the difficulty of a text based on the length of its words and sentences (Flesch, 2007), and is defined as;

$$\text{FKGL} = 0.39\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 11.8\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) - 15.59 \quad (11)$$

The Gunning Fog Index is a readability test used in linguistics to assess the complexity of English writing (Gunning, 1952), and is defined as;

$$\text{GFI} = \frac{0.4 \times \text{Total Words}}{\text{Total Sentences}} + \frac{40 \times \text{Total Complex Words}}{\text{Total Words}} \quad (12)$$

Complexity Analysis We observe and report interesting relationships between query complexity and retrieval and QA performance. We find that the accuracy of the in-context GPT-4o retriever is related to question complexity (Figure 6). The more complex the question in terms of word count, Flesch-Kincaid Grade, or Gunning Fog Index, the lower the QA performance (see Figure 6). In contrast, increasing query complexity improves GPT-4o's retrieval ability, where the additional complexity provides information on source relevancy. However, this relationship does not hold for the fine-tuned UniVL-DR retriever, where question complexity has little effect on retrieval recall or QA accuracy (Figure 7). As such, systems that rely on "in-context" retrieval using GPT-4o are limited by query complexity, but approaches that utilize finetuned retrievers are not.

We note that the opposing relationship between retrieval and QA performance is contrary to the finding that the QA task is heavily dependent on retrieval performance (Figure 2b). The impact of query complexity on task performance is strong enough to overcome this general principle.

A.6 Baseline Models

VLP The VLP transformer model consists of a unified encoder and decoder (Zhou et al., 2020). The VLP architecture is made up of 12 layers of transformer blocks trained according to the BERT bidirectional and the seq2seq objectives where the

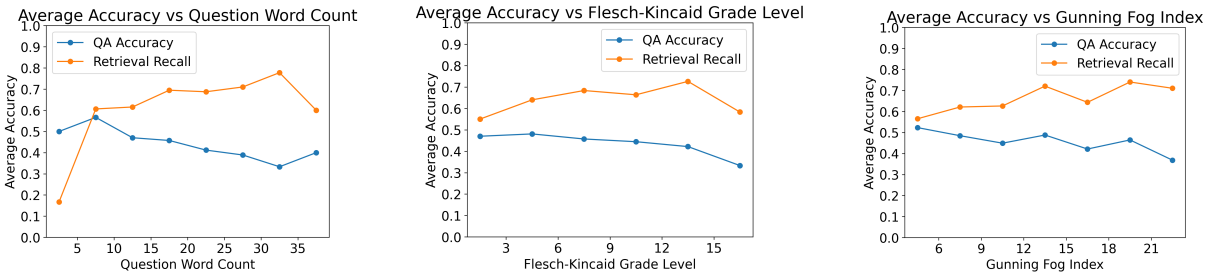


Figure 6: GPT-4o retrieval and QA performance reveal opposite trends with respect to question complexity; GPT-4o retrieval improves with increased complexity, while GPT-4o QA accuracy degrades.

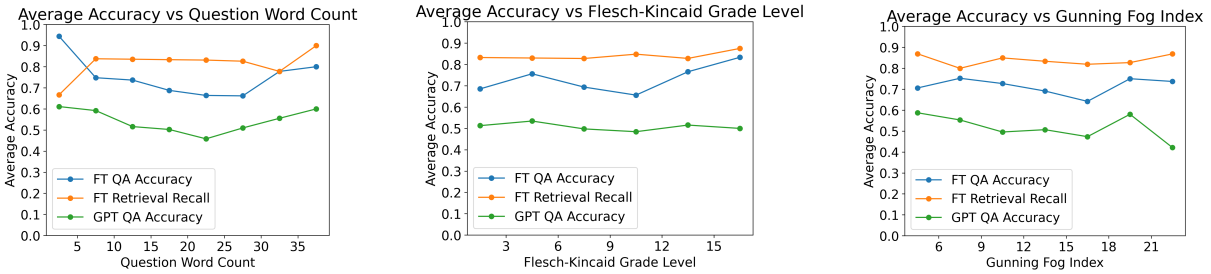


Figure 7: UniVL-DR retriever performance is independent of question complexity. When coupled with this retriever, the effects of question complexity on GPT-4o and MH-VoLTA QA accuracy is minimized.

self-attention module in the transformer block are defined as;

$$A^l = softmax(\frac{Q^T K}{\sqrt{d}} + M)V^T \qquad (13)$$

where $V = W_V^l H^l - 1, Q = W_Q^l H^l - 1, K = W_K^l H^l - 1$. As in (Vaswani et al., 2017), a feed-forward layer (with residual) maps A^l to H^l. The model is trained on image caption pairs, and then finetuned for the VQA task. Finetuning follows by taking the hidden states from the final layer and feeding them to a multi-layer perceptron. The model used has been finetuned twice, once on the VQA dataset (as described by (Yu et al., 2023)), and again on the WebQA dataset.

GIT To contrast with VLP, a pretrained multi-hop VQA model, we use a pre-trained Generative Image-to-Text Transformer (GIT) (Wang et al., 2022). GIT employs a simplified VQA architecture with one encoder for images and one decoder for text. As such, the model is explicitly incapable for multihop VQA between text and images, so it serves as a baseline for pre-trained models that do not utilize image descriptions, and so we concatenate image sources if there are more than one.

GIT is pre-trained using the language modeling task (as opposed to MLM which is used by VLP) where the model learns to predict captions in an auto-regressive manner. For VQA finetuning, the text input is swapped to the query, so that answers are predicted.

BLIP-2 Similar to VLP, the Bootstrapping Language-Image Pre-training model (BLIP) is a unified vision language pre-trained model (Li et al., 2022). It relies on a visual transformer which is less computationally demanding and is pre-trained on over 100 Million image-caption pairs using a contrastive loss (ITC) for image-text contrastive alignment and image-text matching (ITM). In addition to the ITC and ITM losses, the authors introduce an additional Image-grounded text generation (ITG) loss that trains the Q-former encoder to generate texts, given input images as the condition.

GPT-3.5 Turbo Throughout the dataset, a consistent challenge emerges: the model must focus on details, understand them, and accurately respond to questions, even after the provision of positive source images. This challenge has led to the exploration of an image-to-text approach, where the task involves generating descriptive captions for the images. This transforms the problem into a unimodal text retrieval and generation task. Using this method, the SOLAR model has had success on the WebQA task (Liu et al., 2022a). Accordingly, we include a zero-shot oracle baseline, passing queries and image captions to gpt-turbo-3.5 (Brown et al., 2020).

```
yesno_set = {'yes', 'no'}
color_set = {
    'orangebrown', 'spot', 'yellow', 'blue', 'rainbow',
    'ivory', 'brown', 'gray', 'teal', 'bluewhite',
    'orangepurple', 'black', 'white', 'gold', 'redorange',
    'pink', 'blonde', 'tan', 'turquoise', 'grey', 'beige',
    'golden', 'orange', 'bronze', 'maroon', 'purple',
    'bluere', 'red', 'rust', 'violet', 'transparent',
    'yes', 'silver', 'chrome', 'green', 'aqua'
}
shape_set = {
    'globular', 'octogon', 'ring', 'hoop', 'octagon', 'concave',
    'flat', 'wavy', 'shamrock', 'cross', 'cylinder', 'cylindrical',
    'pentagon', 'point', 'pyramidal', 'crescent', 'rectangular',
    'hook', 'tube', 'cone', 'bell', 'spiral', 'ball', 'convex',
    'square', 'arch', 'h', 'cuboid', 'step', 'rectangle', 'dot',
    'oval', 'circle', 'star', 'crosse', 'crest', 'octagonal',
    'cube', 'triangle', 'semicircle', 'domeshape', 'obelisk',
    'corkscrew', 'curve', 'circular', 'xs', 'slope', 'pyramid',
    'round', 'bow', 'straight', 'triangular', 'heart', 'fork',
    'teardrop', 'fold', 'curl', 'spherical', 'diamond', 'keyhole',
    'conical', 'dome', 'sphere', 'bellshaped', 'rounded', 'hexagon',
    'flower', 'globe', 'torus'
}
```

Figure 8: WebQA keyword lists per question category.