RAMQA: A Unified Framework for Retrieval-Augmented Multi-Modal Question Answering

Anonymous ACL submission

Abstract

Multi-modal retrieval-augmented Question Answering (MRAQA), integrating text and images, has gained significant attention in information retrieval (IR) and natural language processing (NLP). Traditional ranking methods rely on small encoder-based language models, which are incompatible with modern decoder-based generative large language models (LLMs) that have advanced various NLP tasks. To bridge this gap, we propose RAMQA, a unified framework combining learning-to-rank methods with generative permutation-enhanced ranking techniques. We first train a pointwise multi-modal ranker using LLaVA as the backbone. Then, we apply instruction tuning to train a LLaMA model for re-ranking the top-k documents using an innovative autoregressive multi-task learning approach. Our generative ranking model generates re-ranked document IDs and specific answers from document candidates in various permutations. Experiments on two MRAQA benchmarks, WebQA and MultiModalQA, show significant improvements over strong baselines, highlighting the effectiveness of our approach. Data and code will be made public once the paper is accepted.

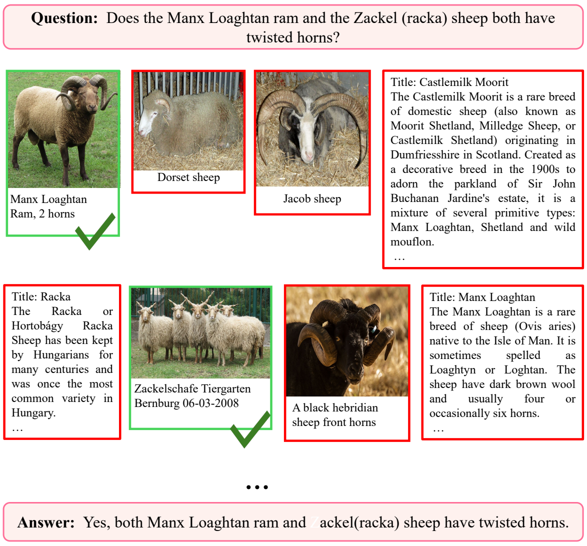

Figure 1: An example in WebQA (Chang et al., 2021), a Multi-modal Open-domain Question-Answering benchmark. This task requires the system to precisely identify critical sources from distractors and use these key sources to infer the answers.

1 Introduction

Multi-modal retrieval-augmented question answering (MRAQA) involves searching and integrating information from diverse modalities such as text and images (Talmor et al., 2021; Chang et al., 2021) (see Figure 1). This capability is crucial for applications requiring comprehensive understanding and reasoning. While powerful generative language models have revolutionized NLP, achieving state-of-the-art results across various tasks (Wu et al., 2024; Touvron et al., 2023; Liu et al., 2023), leveraging these advanced LLMs for information retrieval tasks like MRAQA remains challenging.

Existing MRAQA methods rely on small encoder-based ranking models (Hu et al., 2022b; Yang et al., 2023a,b), which are not fully compatible with modern large generative language models. Although recent generative LLMs trained on massive datasets have dominated NLP tasks, they are typically decoder-only, making it challenging to encode documents into dense representations as encoder-based models do.

Generative retrieval paradigms (Metzler et al., 2021; Tay et al., 2022; Wang et al., 2022b) differ from traditional retrieval methods by directly generating relevant document identifiers for a query. However, applying these methods to multi-modal information retrieval faces challenges: (1) multi-modal documents have aspects not effectively represented by static identifiers; (2) existing multi-modal LLMs are not structured or pretrained to infer across multiple multi-modal documents; (3) LLMs' limited input sequence length hinders ranking many documents in a single run.

To address these challenges, we propose RAMQA, a unified framework combining traditional learning-to-rank methods with generative ranking. First, we train a pointwise multi-modal ranker based on LLaVA (Liu et al., 2023) as a multi-modal data encoder. Second, we employ instruction tuning (Ouyang et al., 2022) to train a LLaMA (Touvron et al., 2023) model to re-rank the top-k documents using a novel autoregressive multi-task learning approach. Before the second-stage retrieval, we unify multi-modal documents into text representations using a zero-shot LLaVA model. This provides context for all candidate documents, reducing the LLM's burden to memorize relationships between queries and document identifiers, making it more efficient than previous methods. Our generative ranking model is trained in a multi-task manner, generating relevant documents and extracting exact answers. To reduce bias from input document sequences, we use permutations of document candidates. We demonstrate the effectiveness of these methods through comprehensive ablation studies.

Experiments on two benchmarks, WebQA (Chang et al., 2021) and MultimodalQA (Talmor et al., 2021), demonstrate significant improvements over strong baselines, highlighting our approach's effectiveness in enhancing multi-modal retrieval-augmented QA systems.[1]

In summary, our contributions are as follows:

- **Unified Framework:** We develop RAMQA, a unified framework for Retrieval-Augmented Multi-modal Question Answering, which combines traditional learning-to-rank methods with generative ranking techniques.
- **Innovative Multi-Stage Process:** We introduce a two-stage approach with a fine-tuned LLaVA for multi-modal pointwise ranking, and a fine-tuned LLaMA for generative re-ranking, enhanced by multi-task learning and document permutation techniques.
- **Comprehensive Evaluation:** We demonstrate the effectiveness of the proposed methods through a thorough ablation study and achieved significant improvements over strong baselines on two benchmark datasets, WebQA and MultimodalQA.

[1] We achieved fourth place on the WebQA leaderboard: https://eval.ai/web/challenges/challenge-page/1255/leaderboard/3168; to our knowledge, the top three works were unpublished at submission time.

2 Related Work

2.1 Multi-Modal Retrieval-Augmented Question Answering

Multi-modal retrieval-augmented question answering (MRAQA) integrates information from various modalities, such as text, images, and tables, to answer complex questions. Benchmark datasets like MultimodalQA (Talmor et al., 2021) and WebQA (Chang et al., 2021) have been developed to address these challenges.

Recent frameworks like MuRAG (Hu et al., 2022b), SKURG (Yang et al., 2023a), and PERQA (Yang et al., 2023b) have made significant strides in MRAQA by integrating text and image data using retrieval and generation techniques. However, these methods primarily rely on encoder-based models and structured knowledge, limiting their ability to fully leverage the capabilities of state-of-the-art multi-modal generative LLMs. Our work addresses this gap by introducing a novel framework that combines traditional ranking with multi-modal generative LLMs, offering a more robust solution for MRAQA.

2.2 Learning-to-Rank

Learning-to-Rank (LTR) techniques optimize item ranking in information retrieval systems based on relevance. These models include pointwise (Cossock and Zhang, 2006; Liu, 2009; Li, 2011; Nogueira and Cho, 2019; Nogueira et al., 2019), pairwise (Freund et al., 2003; Clark et al., 2020; Li et al., 2023a), and listwise (Cao et al., 2007; Ai et al., 2019; Zhang et al., 2018) approaches. The advent of Transformer encoders like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) has significantly enhanced LTR by enabling more accurate relevance scoring.

Recent advancements have explored using large language models (LLMs) in LTR. For example, RankLLaMA (Ma et al., 2024) fine-tuned the LLaMA model, demonstrating that decoder-based LLMs can surpass traditional encoder-based models in ranking tasks. Building on this, we fine-tuned LLaVA (Liu et al., 2023), a multi-modal LLM that combines LLaMA with the CLIP visual encoder ViT-L/14 (Dosovitskiy et al., 2021), creating RankLLaVA, a multi-modal pointwise ranker that enhances ranking performance by leveraging both language and visual data.

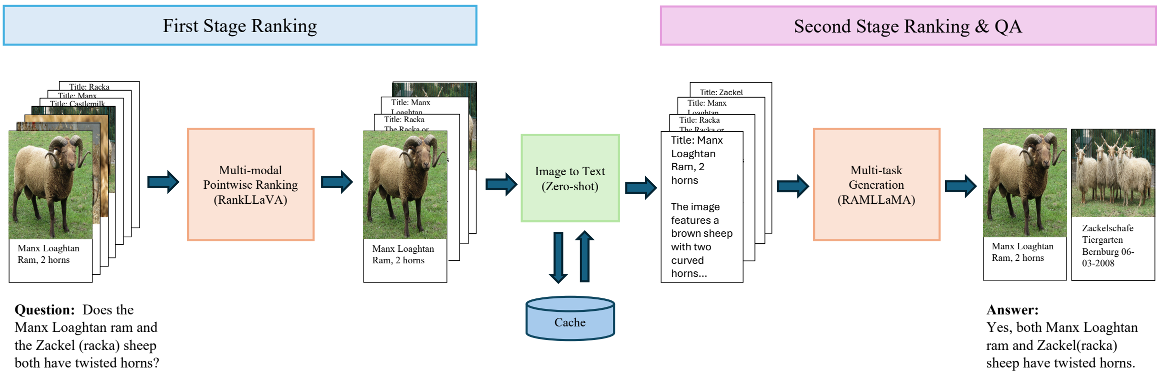

Figure 2: RAMQA Framework Overview. A detailed description of the three main components—RankLLaVA, Data Unification (Image to Text), and RAMLLaMA—is provided in Sections 3.2, 3.3.1, and 3.3.2, respectively.

2.3 Generative Retrieval

Generative retrieval techniques (Tay et al., 2022; Tang et al., 2023; Bevilacqua et al., 2022; Zhang et al., 2024; Li et al., 2023b) represent a shift from traditional retrieval methods by directly generating document identifiers (DocIDs) for a query using generative models. Advances like Differentiable Search Index (DSI) (Tay et al., 2022) and SEAL (Bevilacqua et al., 2022) have introduced more efficient and effective retrieval processes. However, these methods primarily focus on unimodal data and often struggle with integrating multi-modal information.

Our work addresses this limitation by introducing a unified framework that combines multi-modal pointwise learning-to-rank with generative ranking in a two-stage retrieval process, effectively bridging the gap in multi-modal retrieval.

3 Methodology

In this section, we provide a comprehensive description of our proposed framework designed to address multi-modal learning-to-rank and generative retrieval tasks. We start by defining these tasks and then explore the structure and training methodologies of our unified framework, as outlined in Figure 2.

3.1 Preliminaries

3.1.1 Task Definition

Given a question Q and a set of input documents $D = \{d_1, d_2, \ldots, d_n\}$, where n represents the number of documents and each document may be a text with a title or an image with a caption, MRAQA aims to retrieve evidence from D and generate an answer A based on the retrieved evidence. Although the MRAQA task can encompass other document modalities like tables, audio, and video, this paper focuses specifically on text passages and images. Unlike a typical end-to-end multi-stage retrieval pipeline (Yates et al., 2021), which includes a retriever (Karpukhin et al., 2020) to efficiently locate the top-k relevant texts from a corpus, followed by multiple rerankers (Nogueira and Cho, 2019) to refine the retrieved candidates, our approach in this paper centers on the reranking stage. Specifically, we assume the input documents include positive evidence and distractors (hard negatives) from the datasets, rather than the full document corpus.

3.1.2 LLaMA

LLaMA (Touvron et al., 2023) is a large language model based on the Transformer architecture, operating in an auto-regressive, decoder-only manner. With billions of parameters, it is pre-trained on a massive dataset of web content. As a unidirectional model, its attention mechanism only considers the preceding elements in the input sequence to make predictions. Specifically, for a given input sequence $s = [t_1, t_2, \ldots, t_{n-1}]$, the model predicts the next token t_n based solely on the prior tokens. This prediction process is mathematically expressed as $P(t_n|t_1, t_2, \ldots, t_{n-1})$, where P denotes the probability of the next token t_n in the sequence.

3.1.3 LLaVA

LLaVA (Liu et al., 2023) extends the LLaMA model to handle multi-modal inputs, specifically text and images, by incorporating a vision encoder alongside its Transformer-based architecture. LLaVA retains the auto-regressive, decoder-only structure for text generation, while its vision encoder, often based on a pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2021), processes images by extracting a sequence of visual features from different regions (patches) of the image.

These patch-level embeddings are then combined to form a sequence of visual tokens, which are integrated with the text tokens. The resulting multi-modal input sequence $x = [v_1, v_2, \ldots, v_m, t_1, t_2, \ldots, t_{n-1}]$ consists of both visual tokens $v_1, v_2, \ldots, v_m$ from the image and text tokens $t_1, t_2, \ldots, t_{n-1}$ from the query.

This multi-modal sequence is fed into the Transformer model, enabling it to predict the next token t_n based on both visual and textual context. The prediction process is mathematically expressed as: $P(t_n | v_1, v_2, \ldots, v_m, t_1, \ldots, t_{n-1})$, where P represents the probability of the next token t_n, conditioned on both the visual features $v_1, v_2, \ldots, v_m$ and prior tokens. This integration of detailed image features with text allows LLaVA to perform tasks requiring sophisticated reasoning over both visual and textual inputs, such as multi-modal question answering and image captioning.

3.2 RankLLaVA for Multi-modal Pointwise Ranking

Our first-stage ranking model, named RankLLaVA, is trained as a pointwise ranker. This method involves feeding both the query and a candidate document into the model, which then generates a relevance score indicating how well the document matches the query (Nogueira and Cho, 2019). The backbone model is initialized with LLaVA.

Traditionally, pointwise ranking models use bidirectional encoder-only models like BERT, where the [CLS] token is added at the beginning of the input sequence, and its hidden representation is used to represent the entire sequence. In contrast, since LLaVA is unidirectional, we append an end-of-sequence token (</s>) to the input query or document, and the hidden representation of this </s> token is used to represent the input sequence in LLaVA.

RankLLaVA is trained on query-document pairs as detailed in Algorithm 1. To compute the query-document similarity score, we utilize the LLaVA model's image encoder, tokenizer, and decoder as described in (Liu et al., 2023). We process the input through these components to obtain the hidden representations of the tokens. Specifically, we extract the hidden representation of the last token in the sequence from the decoder's last layer. This representation is then passed through a linear layer, and a sigmoid activation function is applied to produce the final similarity score between the query and the document.

Algorithm 1 RankLLaVA Training Procedure

Require: Training dataset $\mathcal{D} = \{(Q_i, d_i, y_i)\}_{i=1}^N$
where Q_i is a textual question, d_i is a multi-modal document with image part d_{i_image} and text part d_{i_text}, and y_i is the ground truth label (1 if d_i is relevant to Q_i, 0 otherwise).
Ensure: Trained RankLLaVA model
for each $(Q_i, d_i, y_i) \in \mathcal{D}$ **do**
  Construct Input:
  Concatenate the document text with an image placeholder:
  d'_i = "`<image>` d_{i_text}"
  Construct the prompt by combining the question and the document:
  prompt = "`Question:` Q_i `Document:` d'_i `</s>`"
  Compute Embeddings:
  Encode the image part using the image encoder:
  $[v_1, v_2, \ldots, v_m] = \text{ImgEncoder}(d_{i\_image})$
  Tokenize the prompt and obtain token embeddings:
  $[t_1, t_2, \ldots, t_n] = \text{Tokenizer(prompt)}$
  Combine image features and token embeddings:
  $\text{emb\_seq} = [v_1, v_2, \ldots, v_m, t_1, t_2, \ldots, t_n]$
  Forward Pass:
  Pass the combined embedding sequence through the decoder:
  $\text{hidden\_states} = \text{Decoder(emb\_seq)}$
  Extract the hidden representation of the last token:
  $h_i = \text{hidden\_states}[-1]$
  Compute the similarity score:
  $\text{Sim}(Q_i, d_i) = \sigma(\text{Linear}(h_i))$
  Compute Loss:
  Compute the cross-entropy loss:
  $\ell_{rank\_i} = -y_i \log(\text{Sim}(Q_i, d_i)) - (1 - y_i) \log(1 - \text{Sim}(Q_i, d_i))$
end for
Update Model Parameters:
Optimize the model parameters to minimize the total loss:
$\mathcal{L} = \sum_{i=1}^N \ell_{rank\_i}$
return Trained RankLLaVA model

3.3 Multi-task Generation

We now introduce the second stage of our framework, which functions as a multi-task generator for second-stage ranking and question answering. This stage is designed to accurately identify the correct documents from the top-k candidates predicted by the first-stage ranker that can assist in answering the question. Simultaneously, it generates the answer based on the identified documents. We experimentally show that this additional objective makes the model's ranking performance more robust.

3.3.1 Data Unification

We begin by unifying data from different modalities by converting images to text using a pre-trained LLaVA model with a customized prompt, following the format used during LLaVA's training. Fig-

Inputs:

- Image:

- Caption:
  Manx Loaghtan Ram, 2 horns

- Question:
  Does the Manx Loaghtan ram and the Zackel (racka) sheep both have twisted horns?

- Prompt:
  "USER: Create a detailed description for the image, <*image*>, with the caption: '{*Caption*}'.

  Your description should provide enough detail to help answer the question: '{*Question*}'.

  However, if the subject mentioned in the caption is not relevant to the question, you can disregard the question when creating your description.

  ASSISTANT:"

Output:

The image features a brown sheep with two curved horns, which are characteristic of a Manx Loaghtan ram.

Figure 3: Zero-shot image description generation for data modality unification.

Algorithm 2 RAMLLaMA Training Procedure

Require: Pre-trained LLaMA model $\mathcal{M}$, training dataset $\mathcal{D} = \{(Q_i, D_i, A_i)\}_{i=1}^{N}$ where Q_i is a question, $D_i = \{d_{i1}, d_{i2}, \ldots, d_{ik}\}$ is the set of top-k candidate documents, and A_i is the ground-truth answer.
Ensure: Fine-tuned RAMLLaMA model $\mathcal{M}'$
for each $(Q_i, D_i, A_i) \in \mathcal{D}$ **do**
  Construct Input Prompt:
  Randomly permute the order of $\{t_{i1}, t_{i2}, \ldots, t_{ik}\}$ to get $\{t'_{i1}, t'_{i2}, \ldots, t'_{ik}\}$
  Create input prompt P_i:
   $P_i \leftarrow$ "Question: Q_i \\Documents:
    [DocID: 1] t'_{i1}
    [DocID: 2] t'_{i2}
    ⋮
    [DocID: k] t'_{ik}"
  Construct Target Output:
  Identify relevant document IDs $R_i \subseteq \{1, 2, \ldots, k\}$ supporting A_i
  Create target output T_i:
   $T_i \leftarrow$ "Relevant Document IDs: R_i \\Answer: A_i"
  Fine-tune LLaMA:
  Optimize $\mathcal{M}$ to minimize loss $\mathcal{L}$ over (P_i, T_i):
   $\mathcal{L} = -\sum_{(P_i,T_i)} \log P_{\mathcal{M}}(T_i \mid P_i)$
end for
return Fine-tuned model $\mathcal{M}'$

ure 3 illustrates an example of our zero-shot image-to-text generation. This transformation serves two key purposes. First, it aligns with the pre-training process of LLMs, which, to our knowledge, have not been pre-trained with multiple images as input. To leverage the existing knowledge of pre-trained multi-modal LLMs and avoid costly re-training, we use LLaVA as a pointwise ranker and single-image description generator rather than as a list-wise multi-modal ranker. Second, transforming images into sentence-level descriptions optimizes input size. By conserving input token capacity, we can include more documents within the LLM's input sequence, ultimately enhancing ranking performance.

3.3.2 RAMLLaMA

Our second-stage ranking and question-answering model, RAMLLaMA (Retrieval-Augmented Multi-task LLaMA), is trained autoregressively using instruction tuning (Ouyang et al., 2022). Given a prompt comprising a question and the top-k unified candidate documents from the first-stage ranking, along with their IDs, the model generates the relevant document IDs and the answer.

To prevent the model from overfitting to the sequence of input documents, we permute the candidate documents five times for each question during training, effectively increasing the training set size fivefold. We demonstrate the effectiveness of this approach in the ablation studies.

The training procedure for RAMLLaMA is detailed in Algorithm 2. Please refer to Appendix A for a training example.

4 Experiments

4.1 Datasets

We conduct experiments on two widely used MRAQA datasets: WebQA (Chang et al., 2021) and MultimodalQA (Talmor et al., 2021). The dataset statistics are presented in Table 1.

4.1.1 WebQA

WebQA (Chang et al., 2021) contains multi-hop, multi-modal question-answer pairs, where each query, typically requiring 1-2 images or text documents, is paired with around 40 multi-modal distractors (hard negatives). Although the input sources are multi-modal, the questions are entirely text-based. Answers are free-form sentences. Evaluation metrics include source retrieval F1 and a QA score, which combines BARTScore-based (Yuan et al., 2021) fluency and relevance (QA-FL) with keyword accuracy (QA-Acc). The overall QA score, a product of QA-FL and QA-Acc, is the key metric for WebQA.

Dataset	Train Image/Text	Dev Image/Text	Test Image/Text
WebQA	18K/17K	2.5K/2.4K	3.4K/4K
MultimodalQA	3.6K/7.5K	371/721	-

Table 1: Overall Statistics of benchmark datasets.

4.1.2 MultimodalQA

MultimodalQA (Talmor et al., 2021) contains multi-modal QA pairs across tables, texts, and images, with 16 question types, 13 of which require cross-modal retrieval and reasoning. As tables are outside the scope of our paper, following (Hu et al., 2022b) we focus on the subset of queries involving only text and image information, specifically selecting questions labeled as 'TextQ' or 'ImageQ'. Each query typically requires 1 image and/or 1 text snippet to answer and is paired with around 20 visual and text distractors. Since test set labels are unavailable, we report RAMQA results on the validation set. The answers are spans or short phrases, and the evaluation metrics are Exact Match (EM) and average F1 as described in (Dua et al., 2019).

4.2 Baselines

We compare RAMQA against SOTA models[2] on WebQA and MultimodalQA in an distractor setting, i.e., the input documents are positives and hard negatives provided by the datasets, rather than the entire document corpus.

4.2.1 AutoRouting

AutoRouting (Talmor et al., 2021) converts multi-modal QA into unimodal QA by using a question-type classifier to identify the modality likely to contain the final answer. It directs the question and input sources to the appropriate QA module (textQ, tableQ, or imageQ) and extracts answer spans using specialized sub-models. AutoRouting employs RoBERTa-large (Liu et al., 2019) for question-type classification as well as textQ and tableQ, while VILBERT-MT (Lu et al., 2019) handles imageQ with image features extracted by Faster R-CNN (Ren et al., 2015).

4.2.2 VLP and VLP + VinVL

Leveraging VinVL (Zhang et al., 2021) for image feature extraction, these transformer-based encoder-decoder models begin by concatenating each document with the question and employing a classifier to estimate the selection probability of each document. The selected documents, along with the question, are then concatenated and fed into the model for answer generation, using a beam search with a size of 5.

4.2.3 MuRAG

MuRAG (Hu et al., 2022b) is pre-trained on a combination of large-scale image-text and text-only corpora. It retrieves the Top-K nearest neighbors from a memory of image-text pairs using a query Q from any modality. The retrieved results are combined with Q and fed into an encoder-decoder for answer generation. During fine-tuning, the question is used as the query Q along with the Top-4 retrieved sources, and a beam search with size 2 is applied. MuRAG uses ViT-large (Dosovitskiy et al., 2021) for image encoding and T5-base (Raffel et al., 2019) for text encoding and answer generation. MuRAG is evaluated only on the text and image subsets of MultimodalQA, excluding the table modality.

4.2.4 SKURG

SKURG (Yang et al., 2023a) integrates evidence features using entity relations and feeds them into a transformer to generate key evidence and answers. It employs OFA-base (Wang et al., 2022a) as the image encoder and BART-base (Lewis et al., 2019) for text and knowledge graph encoding. The BART decoder is then used to generate the relevant document IDs and answers from the encoded documents. The BART-base model is pre-trained on SQuAD2.0 (Rajpurkar et al., 2018). For entity and relation extraction, SKURG uses ELMo-based (Peters et al., 2018) NER (Peters et al., 2017) and OpenNRE (Han et al., 2019), respectively.

4.2.5 PERQA

PERQA (Yang et al., 2023b) is a framework for evidence retrieval and question answering. After pre-processing all images by generating descriptions using image captioning with OFA (Wang et al., 2022a) and object detection with Fast RCNN (Girshick, 2015), it performs iterative pairwise ranking using BERT (Devlin et al., 2019), followed by an extra "evidence refinement" using pointwise reranking with Deberta-large (He et al., 2021). Once the top candidate documents are retrieved, PERQA integrates them into a dialogue format and fine-tunes a multi-modal LLM mPLUG-Owl (Ye et al., 2023), to generate answers based on the retrieved documents and the question.

[2] Note, neither MuRAG and PERQA have published their code.

Model	QA-FL	QA-Acc	QA	Retr-F_1
VLP(Q-only) (2021)	34.9	22.2	13.4	–
VLP (2021)	42.6	36.7	22.6	68.9
VLP + VinVL (2021)	44.2	38.9	24.1	70.9
MuRAG (2022b)	55.7	54.6	36.1	74.6
SKURG (2023a)	55.4	57.1	37.7	88.2
PERQA (2023b)	61.7	63.9	44.4	**89.6**
RAMQA (**ours**)	**64.1**	**66.6**	**48.1**	88.4

Table 2: WebQA official test set results indicated on leaderboard[5] as of August 2024. VLP (Q-only) uses only the question as input for VLP. Bold numbers indicate best and underline the second-best score.

	Text		Image		All
Model	EM	F_1	EM	F_1	EM
Q-only (2021)	15.4	18.4	11.0	15.6	13.8
AutoRouting (2021)	49.5	56.9	37.8	37.8	46.6
MuRAG (2022b)	60.8	67.5	58.2	58.2	60.2
SKURG (2023a)	66.7	72.7	56.1	56.1	64.2
PERQA (2023b)	69.7	74.1	54.7	60.3	62.8
RAMQA (**ours**)	**79.5**	**85.5**	**67.0**	**67.0**	**70.6**

Table 3: MultimodalQA dev-set results on the subset. Q-only denotes using only the question as input for BART-large. Bold numbers indicate best and underline the second-best score.

Model	QA-FL	QA-Acc	QA	Retr-F_1
RAMQA	**63.4**±0.7	**66.2**±0.4	**47.5**±0.6	**88.3**±0.1
w/o Perm	62.4±0.9	64.3±0.6	46.2±0.7	86.2±0.2
Retr-only Gen w/o Perm	-	-	-	84.7±0.2
QA-only Gen	58.6±1.1	60.8±0.8	40.4±0.9	75.4±0.1

Table 4: Ablation study of RAMQA on the WebQA test set. "Perm" refers to generative retrieval with permutation. "Retr-only Gen" and "QA-only Gen" indicate generation with only the retrieval objective and with only the question-answering objective, respectively. The best results for each metric are highlighted in bold. The results are averaged over three runs with different random seeds.

4.3 Implementation Details

The backbone of RankLLaVA is based on the LLaVA-1.5-7B model[3] (Liu et al., 2024). We added a linear layer to project the final layer's end-of-sequence token representation into a scalar, as detailed in section 3.2. Parameter-efficient fine-tuning (PEFT) techniques, including Quantization (Jacob et al., 2017) and low-rank adaptation (LoRA) (Hu et al., 2022a), were used to fine-tune the model on a single NVIDIA A100 80GB GPU with a maximum input sequence length of 2048, a batch size of 8, and gradient accumulation steps of 4. With LoRA, only the linear layer parameters of the LLM were updated, while all other layers, including the visual encoder, were kept frozen.

The backbone of RAMLLaVA is based on the LLaMA-3-70B model[4] (Dubey et al., 2024). Following the instruction tuning approach outlined in Section 3.3.2, we fine-tuned the model using similar PEFT methods. This enabled fine-tuning on a single NVIDIA A100 80GB GPU with a maximum input sequence length of 8192 tokens. We employed a batch size of 2 with 16 gradient accumulation steps. The input data comprised the top 15 ranked documents.

4.4 Main Results

We compare RAMQA against the most relevant methods, including SOTA models.

Table 2 presents the results on WebQA. For the QA score, which is the most critical metric in the WebQA benchmark (described in Section 4.1.1), RAMQA outperforms all baselines, exceeding the SOTA PERQA by 8.3% overall. In terms of Fluency, RAMQA surpasses PERQA by 3.9%, and in Accuracy, it outperforms PERQA by 4.2%. These improvements highlight the high fluency and accuracy of RAMQA's generated answers. In retrieval performance, RAMQA is on par with the SOTA model PERQA. However, unlike PERQA, which relies on textual information retrieval after extensive image processing (generating captions and extracting objects) as described in Section 4.2.5, RAMQA employs true multi-modal IR, directly extracting ranking features from images in the first-stage ranking.

The MultimodalQA results are presented in Table 3. RAMQA significantly outperforms all baselines. For text questions, our model achieves a 14.0% improvement in Exact Match (EM) over the SOTA PERQA. For image questions, the gap is even more pronounced, with a 15.1% improvement over the SOTA MuRAG. Overall, RAMQA surpasses the second-best PERQA by 9.9% in EM.

[3]https://huggingface.co/llava-hf/llava-1.5-7b-hf

[4]https://huggingface.co/meta-llama/Meta-Llama-3-70B

[5]https://eval.ai/web/challenges/challenge-page/1255/leaderboard/3168

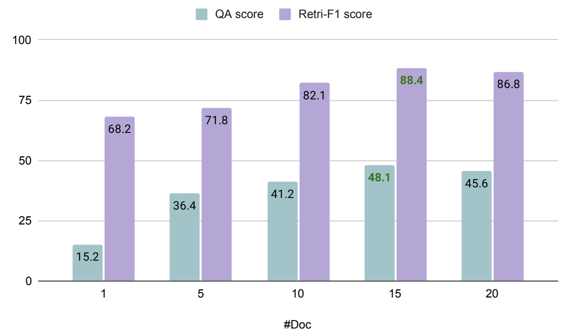

Figure 4: Impact of Input Length on RAMLLaMA Performance on the WebQA test set. The horizontal axis(#Doc) represents the number of candidate documents from RankLLaVA's output included in RAMLLaMA's input prompt during both training and testing. We ensured that the input prompt length did not exceed LLaMA3's limit of 8096 tokens in any of the experiments.

4.5 Ablation Studies

4.5.1 Effectiveness of Permutation-Based Generative Retrieval and Multi-Task Objective Generation.

In this section, we investigate the impact of Permutation-based Generative Retrieval and the multi-task objective generation on the final MRAQA results over the WebQA test set. As shown in Table 4, without the generative retrieval objective, our second-stage generation model achieves an overall QA score of only 40.4. The retrieval F_1 score here reflects the ranking performance of our first-stage model, RankLLaVA. Documents are selected if their binary classification confidence exceeds a specified threshold, determined through tuning on the WebQA development set.

When we introduce the retrieval generation objective during the fine-tuning of our second-stage generative model, both the QA score and retrieval F_1 score see significant improvements. Specifically, the QA score increases by 14.4%, and the retrieval F_1 score rises by 14.3%. This demonstrates the effectiveness of the multi-task objective generation in enhancing the model's generative capabilities.

Furthermore, by introducing the permutation of candidate documents in the training data, the retrieval F_1 score is boosted by an additional 2.6%, and the QA score improves by 4.1%. This indicates that permutation-based generative retrieval not only enhances the model's retrieval performance but also contributes to a better understanding of context, thereby improving overall QA performance.

4.5.2 Impact of Document Count on Ranking Effectiveness.

We investigated the impact of the number of input documents on the performance of our second-stage generation model, RAMLLaMA. Figure 4 illustrates how varying the number of documents in RAMLLaMA's input affects its final performance. We found that increasing the input document count from 1 to 15 improved retrieval performance, suggesting that the model benefited from the higher recall provided by the larger document set. However, increasing the input to 20 documents resulted in a performance decline. This drop is likely due to the lack of additional recall from the top 20 retrieved documents compared to the top 15, combined with the inclusion of less relevant documents, which made it more challenging for RAMLLaMA to process the input effectively, potentially leading to overfitting on irrelevant details.

5 Conclusion

In this paper, we introduced RAMQA, a unified framework for Retrieval-Augmented Multimodal Question Answering that combines traditional learning-to-rank methods with generative permutation-enhanced ranking techniques to address the challenges of multi-modal retrieval-augmented question answering. By leveraging state-of-the-art generative LLMs like LLaVA and LLaMA, RAMQA significantly improves both retrieval accuracy and question-answering performance across diverse data sources, including text and images.

Experiments on two MRAQA benchmarks, WebQA and MultiModalQA, demonstrate significant improvements compared to strong baselines, highlighting the effectiveness of our approach in enhancing multi-modal retrieval-augmented QA systems. The introduction of permutation-based generative retrieval and multi-task learning objectives played a key role in these advancements, contributing to a better understanding of context and more accurate information retrieval.

In conclusion, RAMQA sets a new benchmark in multi-modal question answering, demonstrating the effectiveness of combining traditional and generative approaches. We anticipate that RAMQA and similar models will continue to advance the capabilities of multi-modal information retrieval and generation.

Limitations

While RAMQA demonstrates strong performance and introduces several innovative techniques in multi-modal question answering, it is important to acknowledge its limitations: (1) Dependency on High-Quality Multi-Modal Data: RAMQA's performance is closely tied to the quality and diversity of the multi-modal data available during training. In scenarios where such data is scarce or noisy, the model's ability to accurately retrieve and generate relevant answers may degrade. This limitation is particularly evident in domains where multi-modal datasets are limited or not well-structured. (2) Generalization to Novel Domains: While RAMQA has demonstrated strong results on the WebQA and MultimodalQA datasets, its ability to generalize to entirely new domains or query types remains uncertain. The model may struggle with domain-specific terminology or data formats that were not encountered during training, limiting its applicability in specialized fields. (3) Bias and Ethical Concerns: Despite its sophisticated design, RAMQA is not immune to biases present in the training data. These biases can be reflected in the retrieval and generation processes, leading to outputs that may reinforce existing stereotypes or omit crucial perspectives. Addressing these ethical concerns requires further research and careful consideration.

By recognizing these limitations, we hope to guide future research efforts aimed at overcoming these challenges and improving the robustness, scalability, and ethical integrity of multi-modal question answering systems like RAMQA.

Acknowledgments

Our sincere thanks go out to the anonymous reviewers who took the time to provide constructive comments.

References

Qingyao Ai, Xuanhui Wang, Sebastian Bruch, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2019. Learning groupwise multivariate scoring functions using deep neural networks. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 85–92, New York, NY, USA. Association for Computing Machinery.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. In *Advances in Neural Information Processing Systems*, volume 35, pages 31668–31683. Curran Associates, Inc.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA. Association for Computing Machinery.

Yingshan Chang, Mridu Baldevraj Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2021. Webqa: Multihop and multimodal qa. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16474–16483.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

David Cossock and Tong Zhang. 2006. Subset ranking using regression. In *Proceedings of the 19th Annual Conference on Learning Theory*, COLT'06, page 605–619, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *North American Chapter of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4(null):933–969.

Ross Girshick. 2015. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hexiang (Frank) Hu, Pat Verga, Wenhu Chen, William Weston Cohen, and Xi Chen. 2022b. Murag: Multimodal retrieval-augmented generator.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew G. Howard, Hartwig Adam, and Dmitry Kalenichenko. 2017. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2704–2713.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2023a. Parade: Passage representation aggregation fordocument reranking. *ACM Trans. Inf. Syst.*, 42(2).

H. Li. 2011. *Learning to Rank for Information Retrieval and Natural Language Processing*. Online access: Morgan & Claypool Synthesis Collection Five. Morgan & Claypool Publishers.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023b. Multiview identifiers enhanced generative retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6636–6648, Toronto, Canada. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

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. *ArXiv*, abs/1907.11692.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2019. 12-in-1: Multi-task vision and language representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10434–10443.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page

2421–2425, New York, NY, USA. Association for Computing Machinery.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. *SIGIR Forum*, 55(1).

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *ArXiv*, abs/1901.04085.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy J. Lin. 2019. Multi-stage document ranking with bert. *ArXiv*, abs/1910.14424.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Jiangui Chen, Zuowei Zhu, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. 2023. Semantic-enhanced differentiable search index inspired by learning strategies. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4904–4913, New York, NY, USA. Association for Computing Machinery.

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

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Don Metzler. 2022. Transformer memory as a differentiable search index. In *NeurIPS 2022*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Allen Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2022b. A neural corpus indexer for document retrieval. In *Proceedings of the 36th International Conference on Neural Information Processing*

Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yiqi Wu, Xiaodan Hu, Ziming Fu, Siling Zhou, and Jiangong Li. 2024. Gpt-4o: Visual perception performance of multimodal large language models in piglet activity understanding. *ArXiv*, abs/2406.09781.

Qian Yang, Qian Chen, Wen Wang, Baotian Hu, and Min Zhang. 2023a. Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5223–5234, New York, NY, USA. Association for Computing Machinery.

Shuwen Yang, Anran Wu, Xingjiao Wu, Luwei Xiao, Tianlong Ma, Cheng Jin, and Liang He. 2023b. Progressive evidence refinement for open-domain multimodal retrieval question answering. *ArXiv*, abs/2310.09696.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*.

Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. 2018. Effective user interaction for high-recall retrieval: Less is more. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 187–196, New York, NY, USA. Association for Computing Machinery.

Peitian Zhang, Zheng Liu, Yujia Zhou, Zhicheng Dou, Fangchao Liu, and Zhao Cao. 2024. Generative retrieval via term set generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 458–468, New York, NY, USA. Association for Computing Machinery.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584.

A Training example for RAMLLaMA

For this case study, we qualitatively evaluated the model's capabilities by sampling examples from the WebQA development dataset. We compared the model's outputs against the benchmark's golden answer set and strong baseline models.

Figure 5 illustrates a training example for RAMLLaMA, formatted following Stanford-Alpaca(Taori et al., 2023).

B Case Study

Figure 6 compares the outputs of RAMQA (our model) and MuRAG QA as reported in their paper. RAMQA correctly identified the document annotated in the golden set, which MuRAG missed, and also predicted an additional document. Although this second document isn't in the golden set, it should be considered correct, as it contains crucial information. Even though it occupies only a small portion of the image, it provides the necessary details to answer the question.

Figure 7 presents an instance where RAMQA made a misprediction. Although the answer was absent from the benchmark's golden set, it was factually correct. This highlights a limitation in the benchmark dataset, where the golden set may not fully encompass all valid answers. As a result, strict reliance on standard evaluation metrics may undervalue the model's true performance. Future work should consider expanding the golden answer set and employing more flexible evaluation methods, such as human judgment, for a more accurate assessment.

Figure 8 shows another example of RAMQA's misprediction. This error likely stems from the model's difficulty in distinguishing between similarly named locations and their landmarks. For example, both New York City and Chicago have parks named Washington Square with fountains, but only New York's park features an iconic arch. The model may have focused on the shared name rather than the unique characteristics of each location. This confusion could be due to misaligned image retrieval, a lack of contextual understanding, or overlapping training data. To prevent such errors, the model should prioritize distinctive features, like New York's arch, when processing queries involving similar entities.

Input:

“### Instruction:

First, identify and return all relevant Evidence IDs that can assist in answering the question. Next, attempt to answer the question based on the identified evidence.

Input:

Question: Does the Manx Loaghtan ram and the Zackel (racka) sheep both have twisted horns?

Context 0: Evidence ID: 30203912
-- title: Manx Loaghtan Ram, 2 horns Manx Loaghtan ram with two horns
-- Content: Manx Loaghtan Ram, 2 horns Manx Loaghtan ram with two horns . The image features a brown sheep with two curved horns, which are characteristic of a Manx Loaghtan ram. The horns are curved in a way that resembles the horns of a Gemsbok, a type of antelope.

…

Context 6: Evidence ID: 30060710
-- title: Zackelschafe Tiergarten Bernburg 06-03-2008 Zackel (racka) sheep, white, at Tiergarten Bernburg, Germany
-- Content: Zackelschafe Tiergarten Bernburg 06-03-2008 Zackel (racka) sheep, white, at Tiergarten Bernburg, Germany . The image features a group of five sheep standing next to each other in a grassy field. The horns of the Magyar Racka Juh are curved and pointy, which is a distinctive feature of this breed.

…

Response: ”

Output:

“*** **RETRIEVL RESULT:** 30060710;30203912

*** **ANSWER:** Yes, both Manx Loaghtan ram and Zackel(racka) sheep have twisted horns. <|end_of_text|>”

Figure 5: A Training data example of RAMLLaMA.

C Scientific Artifacts

The licenses for the resources used in this paper are as follows: MultiModalQA (MIT License), WebQA (CC0-1.0 License), LLaVA (Llama 2 Community License), LLaMA (Llama 3 Community License Agreement), and Huggingface Transformers (Apache License 2.0). We have adhered to the intended use of all referenced artifacts in this paper.

Question	In the statue in front of Berlin's Mitte cathedral, is the angel holding a dead or alive boy?		
Models	MuRag	RAMQA	Reference
Document Prediction	 title: Berlin.Dom001	 title: Berlin, Germany (April 2016) - 063  title: Berlin, Germany (April 2016) - 058	 title: Berlin, Germany (April 2016) - 063
Answer Prediction	The answer is holding a dead boy in the statue in front of Berlin's Mitte cathedral.	The angel is holding a dead boy.	In the statue in front of Berlin's Mitte Cathedral, the angel is holding a dead boy.

Figure 6: Prediction Examples of RAMQA vs. MuRAG

Question	What color is the lettering on the sign at Paddy Power and the Admiral casino?	
Models	RAMQA	Reference
Document Prediction	 title: Paddy Power, Belfast, July 2013  title: Admiral casino, King Street, Hammersmith	 title: Paddy Power, King Street, Hammersmith 02  title: Admiral casino, King Street, Hammersmith
Answer Prediction	The lettering on the sign at Paddy Power and the Admiral casino is white.	The color of the lettering on the sign at Paddy Power and the Admiral casino is white.

Figure 7: An Example of RAMQA Mispredictions Caused by Incomplete Document Annotations.

Question	What shape is the fountain near the arch in Washington Square Park?	
Models	RAMQA	Reference
Document Prediction	 title: Washington Square Park Fountain, Chicago	 title: Washington Square Arch-Isabella
Answer Prediction	The fountain near the arch in Washington Square Park is round.	There is a circle shaped fountain near the arch in Washington Square Park.

Figure 8: An Example of RAMQA Misprediction Due to Inability to Distinguish Similarly Named Locations