# ICRet-DVQA: In-Context Retrieval and Efficient Tuning for Document Visual Question Answering

Anonymous ACL submission


## Abstract

Document Visual Question Answering (DVQA) is a task that involves responding to queries based on the content of images. Existing work is limited to locating information within a single page and does not facilitate cross-page question-and-answer interaction. Furthermore, the token length limitation imposed on inputs to the model may lead to truncation of segments pertinent to the answer. In this study, we introduce a simple but effective methodology called ICRet-DVQA, which focuses on retrieval and efficient tuning to address this critical issue effectively. For that, we initially retrieve multiple segments from the document that correlate with the question at hand. Subsequently, we leverage the advanced reasoning abilities of the large language model (LLM), further augmenting its performance through instruction tuning. This approach enables the generation of answers that align with the style of the document labels. The experiments demonstrate that our methodology achieved state-of-the-art or competitive results with both single-page and multi-page documents in various fields.


## 1 Introduction

Document visual question answering possesses considerable practical significance, enabling the swift and accurate extraction of answers from voluminous documents in response to user queries. The challenge of this task lies in comprehending the document content and pinpointing the answers within certain documents solely from the query, such as forms, web pages, newspapers, and various types of information, including large amounts of texts, and complex document layouts (Mathew et al., 2021; Tanaka et al., 2021; Mathew et al., 2022).

Currently, Fine-tuning pre-trained visual document understanding models has produced impressive outcomes in question-answering tasks involving visually rich documents (VRDs) (Xu et al., 2020; Huang et al., 2022; Appalaraju et al., 2023; Yu et al., 2023; Kim et al., 2022; Lee et al., 2023). This suggests that incorporating large-scale, unlabeled training documents in the pre-training phase of document understanding models can significantly enhance their performance in answering questions from VRDs. These approaches invest heavily in comprehending document images. Despite notable advancements, most of them can only accept fixed-length document information, and cannot handle long documents or multi-page documents, there is still a significant journey towards practical application.

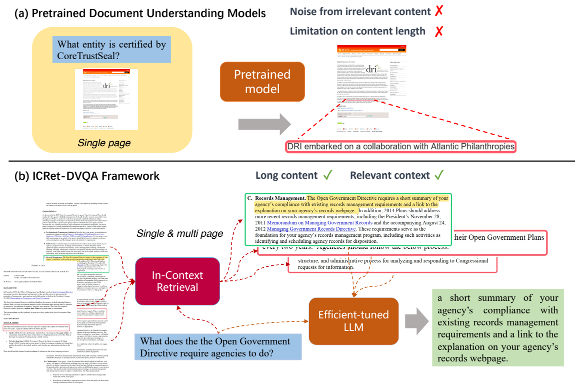



Figure 1: Two approaches to solve the DVQA task. (a) Pretrained models on large-scale documents have questions such as inaccurate answer positioning and limited context length. (b) A framework for contextual retrieval and efficient tuning, which can handle multi-page documents and target answers more accurately. The contents of the boxes all represent the context relevant to the question, where green is the location of the answer.

Large languages models (LLMs), such as GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), PaLM (Chowdhery et al., 2023), develop quickly and have shown remarkable results in various natural language processing (NLP) tasks. Recently, Some methods have attempted to incorporate visual features of documents into LLMs for reasoning (Ye et al., 2023a,b; Zhang et al., 2023c). While certain achievements have been realized, these LLMs fundamentally language-based, exhibit a significant disconnect with visual elements. Consequently, they are incapable of fully comprehending visual information, potentially introducing noise. Additionally, the proficiency of LLMs in managing DVQA tasks remains unexplored. Especially for multi-page documents, the current method still focuses on the single-page document method, that is, initially identify the single page pertinent to the answer, and subsequently utilize the relevant page along with the questions to decode the corresponding answer.

In this paper, we propose ICRet-DVQA, a simple and effective retrieval-augmented and efficient tuning framework for LLMs to perform the DVQA task. Our method comprises three distinct modules: (1) an OCR engine that extracts text from document images. (2) a retrieval module that locates relevant document content based on a given question. (3) a Large Language Model (LLM) which harmonizes the style of data labels and infers appropriate answers based on both the question and the pertinent document content.

Building upon this foundation, we enhanced both the retrieval and instruction-tuning modules. For the retrieval module, a multi-stage retrieval and sorting method was adopted to effectively filter out irrelevant information. Specifically, documents were initially segmented and sorted using a coarse-grained approach, followed by a fine-grained retrieval and selection of highly relevant documents. Ultimately, it is up to the LLM to determine chunks of text based on higher-order semantics. Regarding the instruction-tuning module, we integrated techniques such as prefix tuning (Li and Liang, 2021; Lester et al., 2021; Liu et al., 2021), bias tuning and LORA (Low-Rank Adaptation) (Hedegaard et al., 2022; Hu et al., 2021; Zhang et al., 2023a) to unlock a few parameters of the large model. This approach facilitates efficient tuning and better adaptation to the style of the fine-tuned data.

Experiments conducted on five widely used benchmark datasets (DocVQA (Mathew et al., 2021), VisualMRC (Tanaka et al., 2021), InfographicVQA (Mathew et al., 2022), MP-DocVQA (Tito et al., 2023), DUDE (Van Landeghem et al., 2023)), which encompasses three single-page datasets, along with two multi-page datasets, which cover various fields. Experimental results show that our framework achieves the state-of-art or comparable results compared to previous methods.

Our contributions in this work are four-fold:

- We first propose a simple and effective framework for both single-page and multi-page document question answering across various domains.
- We propose a multi-stage retrieval method to accurately retrieve the relevant context according to the question for DVQA.
- We integrated the efficient tuning approaches that required only 22M training parameters for DVQA.
- ICRet-DVQA achieves state-of-the-art performance in 4 out of 5 document datasets, across multiple domains.

## 2 Related Works

This part mainly reviews the research on document understanding and retrieval augmented generation and highlights the differences between our work and previous work.

### 2.1 Visually Rich Document Understanding (VRDU)

The Visual Reading of Document Understanding (VRDU) is designed to interpret content within document images and is recognized as a challenging task. Existing approaches to VRDU can be broadly categorized based on the use of Optical Character Recognition (OCR) tools. There are two primary types of models: (1)Two-Stage Models Using OCR Tools: These models initially utilize OCR to extract text and layout information from document images. In this approach, specific pre-training tasks are devised to align visual and textual features within a semantic space. For instance, LayoutLMv3 (Huang et al., 2022) incorporates tasks like Masked Image Modeling and Word-Patch Alignment, aiming to harmonize the relationship between textual content and its spatial arrangement in documents. Similarly, UDOP (Tang et al., 2023) introduces tasks such as text-to-layout and layout-to-text conversions, as well as image reconstruction tasks. (2) End-to-End Models Based on Image Features: The pre-training objectives of this method typically involve text recognition tasks akin to OCR, focusing on the nuanced understanding of document images. For example, Donut (Kim et al., 2022) introduces a pretraining task designed to generate all texts present in a document image. Pix2Struct (Lee et al., 2023) pushes the boundaries of traditional document understanding by requiring the model to infer and reconstruct the underlying HTML structure based solely on visual cues from the webpage's layout. Ureader (Ye et al., 2023b) fine-tuned multiple document understanding datasets, including question answering and document summary tasks.

While the aforementioned methods utilize multimodal information from document images, they entail substantial resource consumption for pre-training alignment tasks. Moreover, Most of them can only handle documents with less information on a single page, and they are unable to fully comprehend the nuances of document information, often contending with excessive noise. In our work, recognizing that such tasks are predominantly governed by textual information, we capitalize on the reasoning capabilities of large models. By refining the instruction-tuning method, we achieve a precise generation of answers.

### 2.2 Retrival Augmented Generation (RAG)

RAG significantly enhances the input capabilities of LLMs by integrating retrieved text passages (Lewis et al., 2020; Guu et al., 2020), leading to notable improvements in knowledge-intensive tasks. This enhancement is evident post-fine-tuning and even when used with off-the-shelf LLMs (Ram et al., 2023). Currently, RAG plays a pivotal role in addressing two key challenges associated with LLMs: the hallucination of knowledge and the need for up-to-date information. A more recent advancement (Luo et al., 2023) in the field involves instruction-tuning a Language Model (LM) by appending a fixed number of retrieved passages to the input. This approach is designed to enrich the model's context and understanding by providing additional, relevant information upfront. Furthermore, some methodologies involve jointly pre-training a retriever and an LM, which is then followed by few-shot fine-tuning on specific task datasets.

These methods, while effective in addressing open-domain questions, also encounter significant shortcomings, notably the retrieval of non-relevant information. To address this, we have innovatively applied RAG to the task of DVQA. Our approach enhances the relevance of the information retrieved by implementing a multi-stage retrieval method. This method is meticulously designed to accurately isolate the specific paragraph containing the answer and to eliminate extraneous content.

## 3 ICRet-DVQA

In this section, we introduce a novel framework for executing the task of document visual question answering. We will first overview the procedure of the framework and then elaborate on the technical design of the model architecture.

### 3.1 Overview Framework of ICRet-DVQA

ICRet-DVQA comprises two primary stages: (1) multi-stage retrieval of pertinent document content, and (2) streamlined fine-tuning and inference utilizing the retrieved documents. The process primarily incorporates a vector similarity retrieval module and an LLM, proving to be both straightforward and efficacious.

In the practical execution of document visual question answering, answers frequently reside within one or several segments of the text. Recognizing this, we propose a multi-stage retrieval approach. The primary objective of this method is to incrementally sift through extensive document information to isolate content pertinent to the query. This approach is particularly vital in the context of multi-page or long document visual question answering.

Despite the advanced zero-shot capabilities of LLMs, they lack control over the length and style of generated content. For instance, answers derived from document data can be either extracted or generated. Extraction typically yields brief responses, confined to specific entities; we aim to acquire only the pertinent entity information in such cases. Conversely, for generated answers, detailed reasoning information is desirable. Tailoring tuning instructions for varying data types effectively addresses this issue.

By integrating retrieval and instruction tuning, our framework adeptly manages both the question-answering aspect of large-scale documents and the constraints of large models, thereby producing customized responses tailored to different document data types.

### 3.2 Multi-stage retrieve

Our proposed multi-stage retrieval method effectively filters out content that is most relevant to the answer. Initially, larger text chunks of coarse granularity are selected (for example, setting the length of text chunks to 1024 tokens). Subsequently, smaller text chunks that are most pertinent to the question are retrieved from these larger

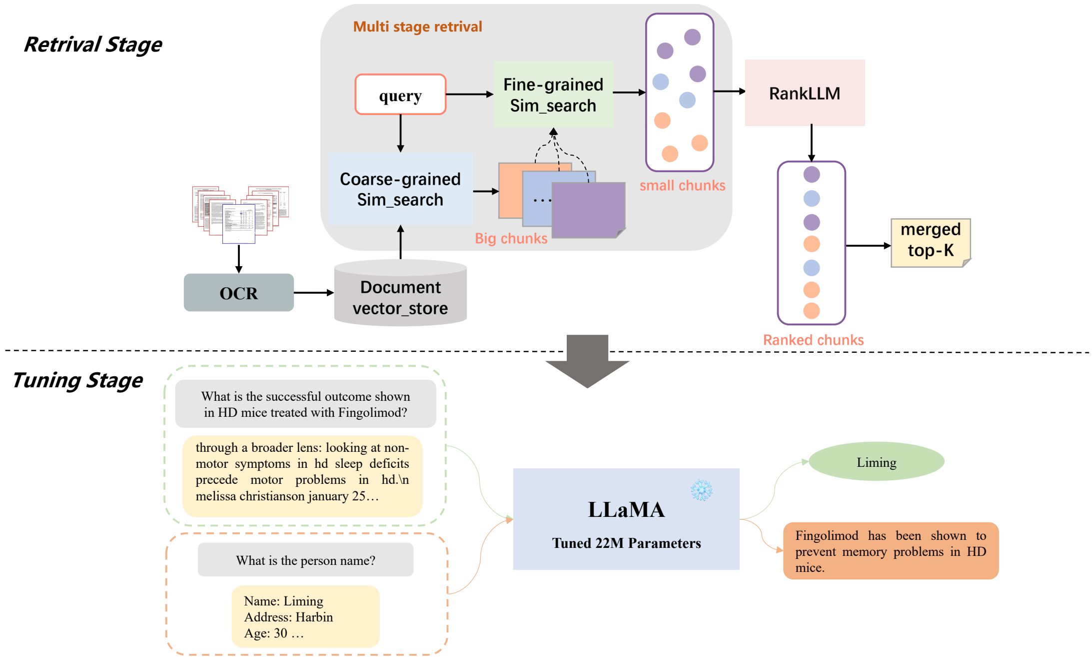



Figure 2: Overview of our ICRet-DVQA framework. ICRet-DVQA consists of three stages: (1) High-performance context retrieval, (2) Efficient parameter instruction tuning and (3) answer inference. In the first stage, highly relevant contextual information is obtained through multi-stage retrieval and large model ranking. Subsequently, an extensive model, aligned with domain-specific data, is developed through efficient parameter fine-tuning. Ultimately, this finely-tuned, comprehensive model is employed for reasoning on the pertinent dataset.

chunks, with the length set to 512 tokens or smaller. All the smaller text chunks are then reordered, and the top-ranked chunks are concatenated to form the context for the model's input. This specific process unfolds as follows.

**Initial Coarse Retrieval**. The set $D_{\text{coarse}}$ consists of larger text chunks retrieved from the extensive document dataset $D$, Semantic similarity matching uses a ready-trained embedding model, and the similarity is calculated with question $Q$, targeting text chunks with a length of 1024 tokens.

$$D_{\text{coarse}} = doc\_vec.match(Q, D) \quad (1)$$

Where $doc\_vec$ indicates the embedding vector of the document associated with the Q. $match$ represents the similarity calculation.

**Fine-grained Retrieval**. Fine-grained retrieval involves retrieving the top-ranked large document chunks to extract the most similar, smaller document chunks from each of these larger chunks.

$$D_{\text{fine}} = \bigcup_{d_i \in D_{coarse}} coarse\_vec.match(Q, d_i) \quad (2)$$

Where, $coarse\_vec$ represents embedding vectors of coarse-grained text chunks, $di \in D_{coarse}$, $Q$ means question or query, targeting text chunks with a length of 512 tokens.

**ReRank**. In conclusion, the smaller text chunks in $D_{fine}$ are ranked based on their relevance to the query $Q$ using a Large Language Model (Pradeep et al., 2023). and the top $k$ ranked chunks are concatenated to form the input context $C$ for the model.

$$C = \text{concat}(\text{LLM\_rank}(D_{fine}, Q, k)) \quad (3)$$

## 3.3 Efficient Tuning

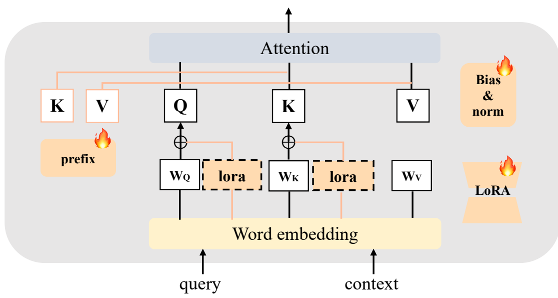



Figure 3: Fusion of different tuning techniques

In this phase, we integrated multiple parameter efficient tuning techniques, such as LoRA (Hedegaard et al., 2022; Hu et al., 2021; Zhang et al., 2023a), prefix tuning (Li and Liang, 2021; Lester et al., 2021; Liu et al., 2021), and bias tuning, while incorporating the zero-initialization attention mechanism from prior methodologies (Zhang et al., 2023b; Gao et al., 2023; Wang et al., 2023), enabling us to achieve sophisticated fine-tuning effects with a minimal number of training parameters.

In the context of LoRA, trainable low-rank matrices are introduced to modify the query and value matrices within the multi-head attention layer. The specific computation is implemented as follows:

Two low-rank matrices, $W_a$ and $W_b$ are initialized. These matrices have dimensions that are significantly smaller than the original query and value matrices, thereby reducing the number of trainable parameters. The original query and value matrices, denoted as $Q$ and $V$ respectively, are modified using the low-rank matrices. This modification is not a direct replacement but an additive update, which can be mathematically represented as:

$$Q' = Q + W_A Q W_B \tag{4}$$

$$V' = V + W_A V W_B \tag{5}$$

Here, $Q'$ and $V'$ represent the updated query and value matrices, respectively.

In prefix tuning, a prefix of length $l$ is strategically positioned preceding the key and value matrices within each multi-head attention layer. This approach effectively equates to adding $l$ additional soft prompt tokens alongside each original token for the computation of similarity measures. The aggregation of these calculations is conducted as follows:

$$\begin{aligned}\text{head}_i = \text{Attn}(xW_q^{(i)}, \text{concat}\left(P_k^{(i)}, CW_k^{(i)}\right), \\ \text{concat}\left(P_v^{(i)}, CW_v^{(i)}\right))\end{aligned} \tag{6}$$

C $\in \mathrm{R}^{\times d}$ indicates the input token sequence. The length is $m$. $W_q^{(i)}$, $W_k^{(i)}$, $W_v^{(i)}$ $\in \mathrm{R}^{d \times d_h}$, $d$ indicates embedding size, $d_h$ means attention hidden size,$P_k^{(i)}$, $P_v^{(i)}$ $\in \mathrm{R}^{l \times d / N_h}$, there are $N_h$ attention heads.

To effectively manage the tasks associated with instruction-following data, same as LLaMA adapterV2 (Gao et al., 2023), we initially unfreeze all normalization layers within LLaMA. For each linear layer in the Transformer, we introduce a bias and a scale factor, both serving as learnable parameters. The input and pre-trained weights of a given linear layer are denoted as $x$ and $W$, respectively.

$$y = W \cdot x \rightarrow y = s \cdot (W \cdot x + b), \tag{7}$$

$$\text{where } b = \text{Init}(0), s = \text{Init}(1). \tag{8}$$

We initialize the bias and scale factors with zeros and ones, respectively, to stabilize the training process at the early stages.

# 4 Experiments

This section describes the relevant datasets and technical implementation details and compares them with other methods. We will present our experimental results and findings.

## 4.1 Experimental Setup

**Datasets.** We experiment on five widely used DVQA datasets, including three datasets focused on single-page DVQA and two datasets dedicated to multi-page DVQA. Here is a brief introduction to these datasets: DocVQA dataset (Mathew et al., 2021) comprises a substantial collection of scanned and handwritten documents, where the answers, typically one or more entities, are extracted from the text. In contrast, the visualMRC dataset (Tanaka et al., 2021), sourced from web page data rich in content, requires answers to be deduced from the original text rather than being directly extracted. The InfographicVQA dataset (Mathew et al., 2022) features documents with a range of chart information and complex layouts, often necessitating inferential calculations and presenting significant challenges. MP-DocVQA (Tito et al., 2023) and DUDE (Van Landeghem et al., 2023) are datasets designed for multi-page document visual question answering, where answers are confined to the specific page, it's crucial to first accurately locate the page containing the answer before attempting to provide a response.

**Compared Methods.** We evaluated three types of pre-training models based on their use of different modal information. The first type includes models like BERT (Kenton and Toutanova, 2019) and T5 (Raffel et al., 2020), which utilize plain text information. The second type encompasses models such as LayoutLMv3 (Huang et al., 2022), UDOP (Tang et al., 2023) and HiVT5 (Tito et al., 2023), leveraging a combination of text, image, and layout information. Among them, HiVT5 is a model specifically for multi-page document question answering. Both of the two methods employ OCR tools for text extraction. In these models, OCR recognition errors can significantly impact performance. However, with LLMs, the construction

| Model | Train Param | Modality | Single-page DVQA | | | Multi-page DVQA | |
|---|---|---|---|---|---|---|---|
| | | | DocVQA | VisualMRC | InfoVQA | MP-DocVQA | DUDE |
| BERT | 334M | T | 67.5 | - | - | 34.8 | 25.5 |
| T5 | 223M | T | 70.4 | 318.6 | 36.7 | 41.8 | 38.7 |
| HiVT5 | 316M | T+L+V | - | - | - | 62.0 | 23.1 |
| LayoutLMv3 | 125M | T+L+V | 78.7 | - | 45.1 | 42.7 | 20.3 |
| UDOP | 794M | T+L+V | **84.7** | - | 47.4 | - | - |
| Donut | 176M | V | 67.5 | 93.91 | 11.6 | - | - |
| Pix2struct | 1.3B | V | 76.6 | - | 40.0 | - | - |
| Ureader | 86M | V | 65.4 | 221.7 | 42.2 | - | - |
| ICRet-DVQA (Ours) | 22M | T | 78.9 | **378.0** | **52.4** | **65.2** | **61.3** |

Table 1: Results of comparing ICRet-DVQA with existing pre-trained VDU models fine-tuned with five different categories of document visual question answering datasets. Following previous works, DocVQA, InfoVQA, MP-DocVQA, and DUDE are evaluated by ANLS, and VisualMRC is measured by CIDEr. Modality T, L, and V denote text, layout, and vision.

of sentence semantics is robust enough that minor word errors do not substantially alter the overall sentence meaning and can often be auto-corrected. Thus, the OCR tool becomes a powerful asset for LLMs.

The third type, represented by models like Donut (Kim et al., 2022), Pix2struct (Lee et al., 2023), and UReader (Ye et al., 2023b), relies on pure image data with high resolution. Document images, distinct from actual scene images, are predominantly text-based and carry richer semantic content. To maximize the reasoning capabilities of LLMs, we use OCR to extract plain text information. By selecting content relevant to the query, we aim to achieve a comprehensive understanding of the document.

**Implementation Details.** In our experiments, the effectiveness of the retrieval module is important, leading us to compare several embedding models [1], including e5-large, instructor-large, and bge-large. Given the extensive volume of data involved in our testing, we ultimately selected the open-source bge-large as our retrieval model due to its robust performance. For the retrieval framework, we utilized Langchain [2], known for its efficiency in handling complex retrieval tasks. Furthermore, we employed open-source RankVicuna (Pradeep et al., 2023), as our ranking model. This choice was driven by its advanced language understanding capabilities.

During the instruction-tuning experiment, we opted for LLaMA 7B as the backbone and utilized a single NVIDIA Tesla A100 80G GPU to train each dataset for 10 epochs. The batch size was set to 8, with a learning rate of 5e-5 and a weight decay of 0.02. We also set the prefix length to 10, applying insertions exclusively in the last 30 layers of the model. For LoRA, we set the rank r to 8.

[1]https://huggingface.co/spaces/mteb/leaderboard
[2]https://www.langchain.com

### 4.2 Main Results

Table 1 presents the performance comparison of ICRet-DVQA with eight models on five datasets. Besides DocVQA, ICRet-DVQA demonstrated superior results in the rest datasets. Although UDOP is currently the most advanced effect, its training cost is relatively high. On the contrary, we have maximized our proficiency in plain text analysis, and we employed the fewest training parameters relative to other methodologies, outperforming several multimodal approaches, including LayoutLMv3. For the challenging datasets, VisualMRC and InfoVQA, our methods demonstrate a substantial lead, outperforming competitors by 60 and 5 points, respectively, thereby achieving the most advanced results to date. Furthermore, while HiVT5 shows improved performance on MP-DocVQA, it falls short on DUDE. In contrast, our method consistently achieves state-of-the-art results across these datasets, which goes far beyond the baseline, such as MP-DocVQA exceeding HiVT5 by 3.2 points and DUDE exceeding T5 by 23.4 percentage points, with solid performance on multi-page documents. More detailed analysis can be found in Appendix A.

| | DocVQA | InfoVQA | VisualMRC |
|---|---|---|---|
| ICRet-DVQA | 78.9 | 52.4 | 378.0 |
| w/o RE | 75.0 | 51.6 | 328.5 |
| w/o PB | 77.7 | 51.2 | 289.6 |
| w/o LO | 75.6 | 51.5 | 295.6 |

Table 2: The effect of different components in ICRet-DVQA. RE means Retrieve. PB means Prefix and Bias tuning. LO means LoRA.

| Retrieval Strategy | MP-DocVQA | DUDE |
|---|---|---|
| SR | 59.5 | 58.4 |
| MR2 | 65.2 | 61.3 |

Table 3: The effect of different retrieval strategies. SR indicates Single-stage Retrieval, MR2 means Multi-stage Retrieval then Ranked by LLM.

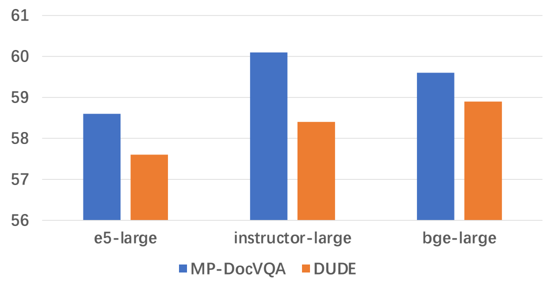



Figure 4: The effect of different text embedding models

### 4.3 Ablation Study

In this section, we conduct a detailed analysis of ICRet-DVQA and its components. To evaluate the impact of individual components on model efficacy, we executed ablation studies across three categories of single-page documents, methodically omitting one module at a time. It is noteworthy that VisualMRC is measured by CIDEr (Vedantam et al., 2015), and other datasets are evaluated by ANLS (Biten et al., 2019)

**Effect of Retrieval Module.** The majority of documents exceed 512 tokens, leading most models to employ a truncation strategy that risks omitting crucial answer segments. However, our retrieval approach accommodates the input length constraints while retaining the most pertinent sections, significantly enhancing model performance. Table 2 illustrates that the exclusion of the retrieval module results in diminished performance. Notably, During the experiment, we found through the experimental results that the retrieval module exerts the most substantial impact, underscoring the efficacy of our retrieval technique across diverse test sets.

**Effect of Prefix & Bias Tuning and LoRA.** Prefix & Bias tuning and LoRA are orthogonal approaches that do not interfere with each other. Prefix tuning allows for the injection of distinct information across various layers of the model, while Lora can further minimize the number of parameters required, without compromising on model performance. According to the experimental results in Table 2, it is found that Prefix & Bias tuning and loRA have different performances on different datasets, and better results can be obtained by combining the two.

**Effect of Retrieval Strategy.** The Table 3 presents a comparative analysis between two approaches: employing the retrieval model for directly fetching the context of relevant text chunks, and a multi-stage retrieval process followed by ranking text chunks. This comparison is specifically focused on multi-page documents, offering a more illustrative perspective. The findings indicate that our adopted strategy is markedly more effective, particularly in the context of multi-page documents, where it has led to significant improvements.

**Effect of Text Embedding.** The success of our retrieval process is heavily reliant on the quality of text embedding, emphasizing the need for an embedding model with superior semantic matching capabilities. In order to highlight the performance of the text embedding, we choose the text chunk of 256 tokens to conduct experiments and compare it to three open-source embedding models, including e5-large, instructor-large, and bge-large. As indicated in Figure 4, The comparative evaluation revealed that bge-large consistently outperforms the others in terms of average performance. Therefore, we ultimately selected bge-large as the retrieval model for all datasets.

### 4.4 Qualitative Results

Figure 5 shows some qualitative results produced by our ICRet-DVQA on different types of documents. ICRet-DVQA is adept at extracting answers from documents with complex layouts (case a) and performing reasoning based on the document's content (case b). Furthermore, in multi-page documents, which typically encompass extensive textual information, ICRet-DVQA can precisely pinpoint

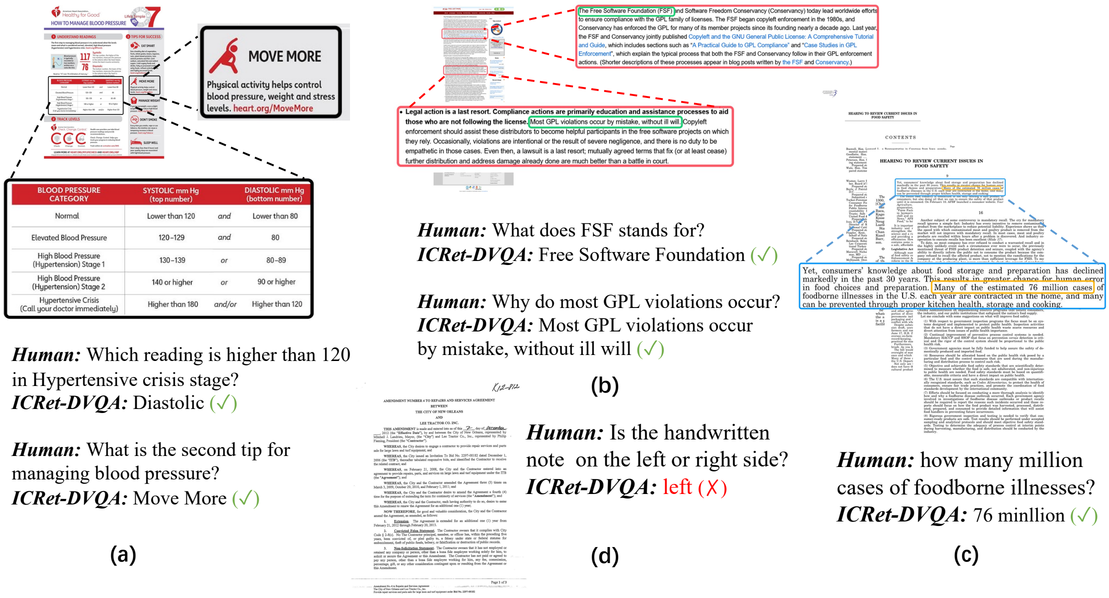



Figure 5: Qualitative results of ICRet-DVQA. Crucial regions are enlarged for clearer visualization.

the location of an answer, as exemplified in case c. Nonetheless, the use of OCR tools in ICRet-DVQA introduces certain constraints. Specifically, text extraction occurs sequentially from top left to bottom right, resulting in the omission of layout and visual information of the document content. Therefore, the questions such as case d cannot be answered. This is what we will try to solve in the future. More qualitative results can be found in Appendix B.

## 5 Conclusion

In this study, we introduce ICRet-DVQA, a comprehensive and versatile framework designed for DVQA tasks applicable across diverse disciplines. This framework is adept at processing extensive documents, including those that span multiple pages. Moreover, we have enhanced the existing retrieval module by implementing a multi-stage retrieval strategy, thereby enabling the precise extraction of pertinent context. Additionally, we have developed an innovative instruction-tuning approach that amalgamates various techniques while minimizing parameter usage to optimize performance. ICRet-DVQA has demonstrated state-of-the-art performance on a wide array of datasets, encompassing both single-page and multi-page documents across various fields, surpassing the capabilities of existing multimodal models.

## 6 Limitations

Our experimental results affirm the efficacy of the ICRet-DVQA framework in processing text-centric document images. However, a limitation of the current iteration of ICRet-DVQA is its omission of layout and image information within documents, which hampers its ability to adequately address text-specific characteristics such as image, style, and positioning. Future work will focus on developing more robust methods to compensate for these deficiencies. A significant challenge we currently face is how to integrate image and layout information into the framework without compromising the efficiency and capacity of larger models. Furthermore, while our framework enhances the accuracy of retrieval, it reduces efficiency. Moving forward, we plan to explore more methods to further optimize the retrieval performance.

## 7 Ethics Statement

This paper proposes a general framework for document visual question answering. We worked within the purview of acceptable privacy practices and strictly followed the data usage policy. In all the experiments, we use public datasets and consist of their intended use. We neither introduce any social/ethical bias to the model nor amplify any bias in the data, so we do not foresee any direct social consequences or ethical issues.

## References


Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. 2023. Docformerv2: Local features for document understanding. *arXiv preprint arXiv:2306.01733*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Lukas Hedegaard, Aman Alok, Juby Jose, and Alexandros Iosifidis. 2022. Structured pruning adapters. *arXiv preprint arXiv:2211.10155*.

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. *arXiv preprint arXiv:2106.09685*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv: 2309.15088*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

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*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023b. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. Structextv2: Masked visual-textual prediction for document image pre-training. *arXiv preprint arXiv:2303.00289*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023c. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

# A Additional Analysis on Performance

## A.1 Compared with more methods

Table 4 enumerates the current state-of-the-art models in terms of performance, and the results indicate that our approach maintains its status as the leading state-of-the-art on all datasets, except for DocVQA. Upon analyzing the error instances in the DocVQA dataset, it was observed that most errors arise from the failure to recognize the layout and visual elements of the document, which is a known limitation of OCR. Moving forward, we aim to delve deeper into the multi-modal aspects of the framework to address and mitigate this issue.

## A.2 Further analysis of ICRet-DVQA

The effectiveness of the retrieval is influenced by multiple factors. Besides the inherent performance of the text embedding and the size of the segmented text chunk, the retrieval strategy employed is also a critical component. We contend that relying solely on the vector similarity of embeddings is insufficient. The semantic correlation between the query and its context is tenuous, necessitating an approach that stems from the text itself. To address this, we have incorporated LLM ranking, a strategy aimed at further enhancing retrieval performance.

We conducted a statistical analysis on the test set of each dataset, employing a threshold of 0.6 to differentiate between positive and negative samples. Samples exceeding the threshold were categorized as positive, while those below it were deemed negative. Subsequently, we calculated the proportion of positive samples within the test set to evaluate the alignment between the retrieved context and the posed question. The algorithm implemented for this analysis is outlined below:

Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of scores for each sample in the test set, where $n$ is the total number of samples.

Define a function classify$(s)$ as:

$$\text{classify}(s) = \begin{cases} 1 & \text{if } s > 0.6 \\ 0 & \text{if } s \leq 0.6 \end{cases} \tag{9}$$

where 1 represents a positive sample and 0 represents a negative sample.

The proportion of positive samples $P$ in the test set is calculated as:

$$P = \frac{\sum_{i=1}^{n} \text{classify}(s_i)}{n} \tag{10}$$

This formula represents the process of classifying each sample based on the threshold, and then computing the ratio of positive samples to the total number of samples in the test set.

As depicted in Figure 6, the test set of each dataset corresponds to the results of similarity matching across various retrieval strategies, including single-stage retrieval and our multi-stage retrieval method. It is evident that the application of our strategy enhances the matching effectiveness. It is worth noting that in the InfoVQA dataset, single-stage retrieval slightly outperforms multi-stage retrieval. This is attributed to the dataset's composition, which includes not only non-lengthy document data but also various structured discrete data. Consequently, relevant content is accurately matched using smaller text chunks.

While we employed OCR to extract text information and organize the sequence of text from the upper left to the lower right corner, the potent capabilities of the large-scale model enable it to respond to queries involving structured data, as illustrated in Figure 7. However, as outlined in the limitations section, the current iteration of ICRet-DVQA does not possess the capability to recognize visual information.

Furthermore, ICRet-DVQA also has the potential to deal with more complex logic problems such as multi-hop question answering (Yang et al., 2018), which is a unique advantage over current methods.

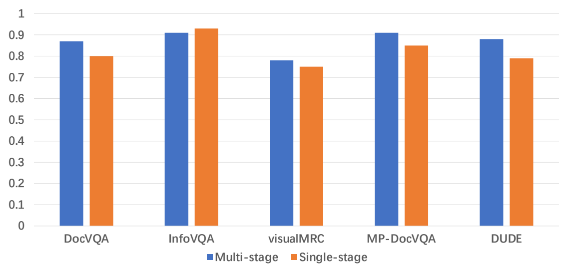



Figure 6: Similarity results of different data sets on different retrieval strategies.

# B More Qualitative Results

## B.1 Results from different datasets

ICRet-DVQA demonstrates the ability to accurately identify corresponding values in charts with complex layouts, which likely reflects the inherent robustness of the large-scale model. As illustrated in Figure 8 a, the model successfully interprets a web page containing multiple charts. Despite

| DocVQA | InfoVQA | VisualMRC | MP-DocVQA | DUDE |
|---|---|---|---|---|
| 84.7(UDOP) | 47.4(UDOP) | 364.2(LayoutT5) | 62.0(HiVT5) | 50.0(DocGptVQA) |
| 78.9 | 52.4 | 378.0 | 65.2 | 61.3 |

Table 4: Performance comparison between ICRet-DVQA and state-of-the-art methods.

the OCR-identified sequence lacking explicit document layout information, the model consistently provides accurate answers.

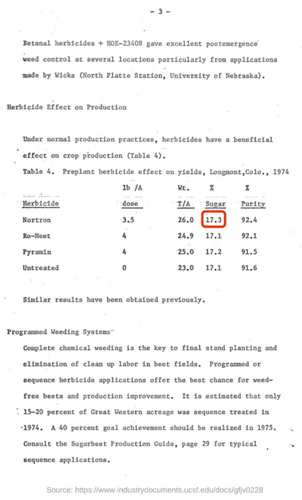
Concurrently, it is noteworthy that despite the expansion in the acceptable character length for LLMs, certain limitations persist. These models often exhibit sensitivity to the beginning and end of sentences and may selectively overlook some contextual information in the middle. Therefore, the retrieval of pertinent information and the condensation of context remain critical steps in the process. As shown in Figure 8 b, ICRet-DVQA first locates the context related to the question, thereby improving the accuracy of the answer.

### B.2 Examples of different text embedding

Figures 9 and 10 illustrate the distinct performances of three different embedding models. Here we set up a single-stage retrieval with a text chunk length of 256 tokens. It is observed that bge-large demonstrates the highest accuracy in context localization, followed by instructor-large, with e5-large ranking last. The efficacy of the embedding model is crucial in the retrieval module, showing a positive correlation with the overall performance in document-based question answering.

### B.3 Examples of different retrieval methods

Figures 11 and 12 present a comparison between the outcomes of single-step retrieval and our multi-stage retrieval process with the LLM ranking method. The examples clearly indicate that the content retrieved through our method exhibits greater accuracy. To preserve the coherence of the context, selecting excessively short text chunks is inadvisable. This has been empirically validated: text chunks comprising 512 tokens demonstrate greater similarity compared to those with 256 tokens. By applying a filter through LLM, we can obtain a context of the highest quality.



\- 3 -

Betanal herbicides + HOE-23408 gave excellent postemergence weed control at several locations particularly from applications made by Wicks (North Platte Station, University of Nebraska).

Herbicide Effect on Production

Under normal production practices, herbicides have a beneficial effect on crop production (Table 4).

Table 4. Preplant herbicide effect on yields, Longmont,Colo., 1974

| Herbicide | lb /A dose | Wt. T/A | % Sugar | % Purity |
|---|---|---|---|---|
| Nortron | 3.5 | 26.0 | 17.3 | 92.4 |
| Ro-Neet | 4 | 24.9 | 17.1 | 92.1 |
| Pyramin | 4 | 25.0 | 17.2 | 91.5 |
| Untreated | 0 | 23.0 | 17.1 | 91.6 |

Similar results have been obtained previously.

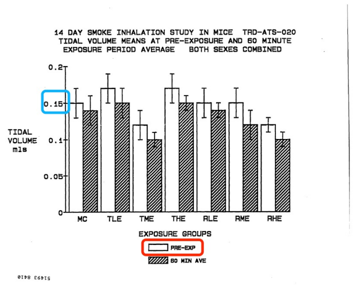
Programmed Weeding Systems

Complete chemical weeding is the key to final stand planting and elimination of clean up labor in beet fields. Programmed or sequence herbicide applications offer the best chance for weed-free beets and production improvement. It is estimated that only 15-20 percent of Great Western acreage was sequence treated in 1974. A 40 percent goal achievement should be realized in 1975. Consult the Sugarbeet Production Guide, page 29 for typical sequence applications.

Source: https://www.industrydocuments.ucsf.edu/docs/gfjv0228

**Human:** what is the % of sugar in nortron?
**ICRet-DVQA:** 17.3 (✓)

**(a)**

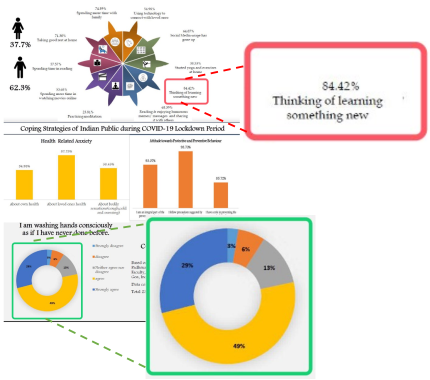



**Human:** What does the unshaded bar represent?
**ICRet-DVQA:** EXPOSURE GROUPS (✗)

**Human:** What is the pre-exposure tidal volume (mls) for MC?
**ICRet-DVQA:** 0.15 (✓)

**(b)**

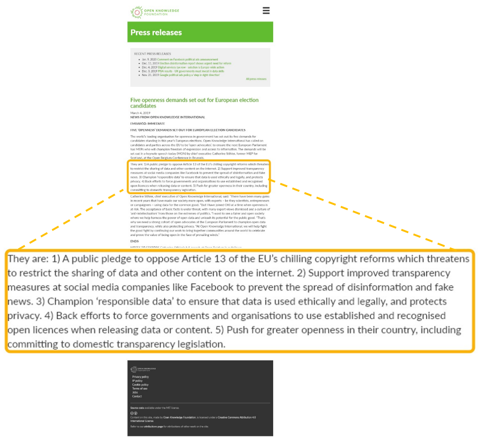
Figure 7: In the case of the DocVQA dataset, the section encircled in the box represents the answer.




**Human:** What percentage of people invest their time to study something new?
**ICRet-DVQA:** 84.42% (✓)

**Human:** What percentage of people agree with washing their hands often ever than before?
**ICRet-DVQA:** 49 (✓)

**(a)**

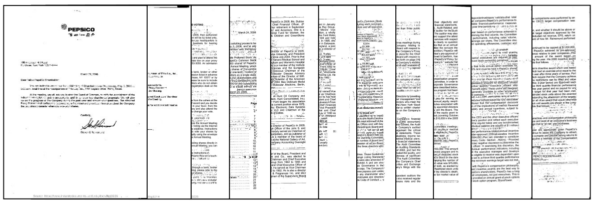



**Human:** What is the third demand in Catherine Stihler's keynote speech?

**ICRet-DVQA:** The third demand is to champion responsible data to ensure that data is used ethically and legally, and protects privacy. (✓)

**(b)**

Figure 8: In the case of the InfoVQA and VisualMRC datasets, Crucial regions are enlarged for clearer visualization.



***Bge-large:***

***Context retrieved***: by the Please vote your shar es promptly through any of the méans describe d on the proxy card. Board for election at …… The Governance Committee shall review the de sired experience, account the current Board wil l generally look for individuals who have displ ayed high ethical Board members and the speci fic needs of the Corporation and the standards, Larry D. Thompson Secretary A-1 26 integrity, and …… best interests under of the Enrollment " links and follow the instructions provided. to circumstances and (3) in a manner he or she bel ieves to be ……
***Question :*** who is the secretary of the compan y?
***Answer :*** Larry D. Thompson
***Real answer:*** Larry D. Thompson

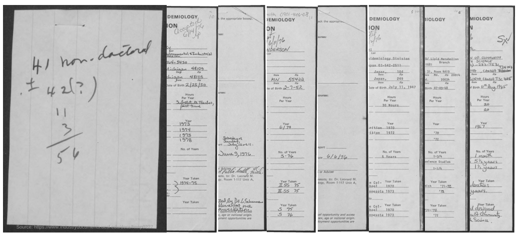
***instructor-large:***

***Context retrieved:*** Our business and affairs ar e overseen by our 34,869 shares; James J. Sch iro, 10,377 shares; Franklin A. Thomas, 39,24 4 shares; Cynthia Business Corporation- Act a nd our By-Laws. Members of the Board of :Di rectors are kept informed of the M. Trudell, 26, 866 shares; Daniel Vasella, 10,417 shares; Mi chael D. White …… Prior various engineering and manufacturing supervisory positions. In 1 995, Có. as' chemical process engineer. In 198 1, she joined GM and held she to joining Peps iCo, Ms. Nooyi spent four years as Senior Vic e President of Strategy, Planning and ……
***Question :*** who is the secretary of the compan y?
***Answer :*** M. Trudell
***Real answer:*** Larry D. Thompson

***e5-large-v2:***

***Context retrieved:*** to solicit proxies from their principals and will pay the brokers and bank bri s Oversee and interact, with senior, manageme nt with respect to key aspects succession, of the nominees their expenses for the solicitation. E mployees of PepsiCo may also solicit proxies. They will business `including strategic plannin g, management development and not.receive an y additional pay for the solicitation ……
It is Board policy that executive officers and ot her members of senior management who report directly to the CEO be present at Board meetin gs at the invitation of the ……
***Question :*** who is the secretary of the compan y?
***Answer :*** William J. Darby
***Real answer:*** Larry D. Thompson

Figure 9: Case comparison of different embedding models. The bge-large is superior to instructor-large and e5-large.



***Bge-large:***

***Context retrieved***: and Abnormalities M-W-F 1 7. Please make the following room reservation f or me: / Middlebrook Hall (Two room, twin bed s) Arrive 6/20 Depart 7/10 in a 18. Signature of Applicant Mah D. Andow Date 3/25 19. I appro ve of this application Pr.s. Pasternock Departme nt Chairman or Adviser Send this form with che ck for $25.00, made payable to the University o f Minnesota, to: Dr. Leonard M. Schuman …… Harmard Course Title School m Public Hours H ealth School Year Taken 15. Have you had any graduate level courses in epidemiology?
***Question :*** What is the arrival date?
***Answer :*** 6/20
***Real answer:*** 6/20 or 6-20

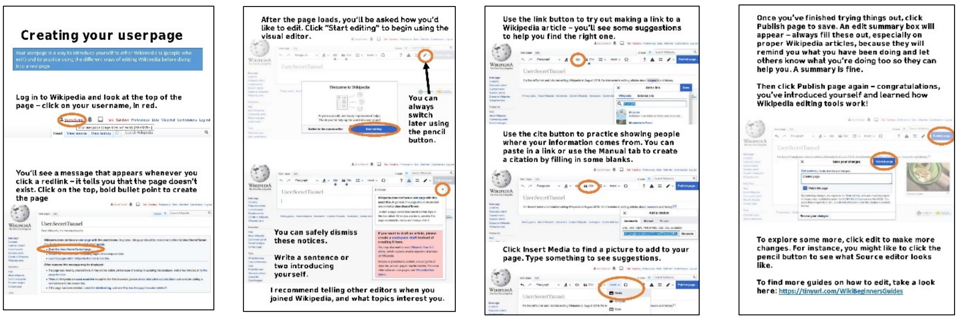
***instructor-large:***

***Context retrieved:*** Please make the following r oom reservation for me: XX Middlebrook Hall (Two in a room, twin beds) Arrive 6-20 Depart 7-10 18. Signature of Applicant Lusly SRisbor d Date 4/25/26 19. I approve of this applicatio n Cary H. Spincy Department (Assist prf Chajr man Adviser) Advisér or Send this form with c heck for $25.00, made payable to the Universit y of Minnesota, to: Dr. Leonard M. Schuman, Program Director, Epidemiology Summer Sess ion, Division of Epidemiology, Room 1-117 U nit A, ……
***Question :*** What is the arrival date?
***Answer :*** 6-20
***Real answer:*** 6/20 or 6-20

***e5-large-v2:***

***Context retrieved:*** Office Address No. Street Cit y State Zip 7. Home Address 323 4th ST S.E. M NCLS MIN 55414 No. Street City State Zip 8. H ome Phone 3381936 9. Sex Female 10. Date of Birth 4/2152 11. Teaching Activities Types of H ours School or Agency Subject Students Per Yea r 12. College, University, and Professional Educ ation School Major Degree Year LALUPENCE UNIV BIOLOGY BA 1974 13. Professional Exp erience (Start with Present Position) Employer P osition Duty No. of Years UNIU OF wis. LAB T ECH ANTIBICTIC RESEARCH 2 14. ……
***Question :*** What is the arrival date?
***Answer :*** 7/27/76
***Real answer:*** 6/20 or 6-20

Figure 10: Case comparison of different embedding models. The bge-large is superior to instructor-large and e5-large. Bge-large and instructor-large are better than e5-large.



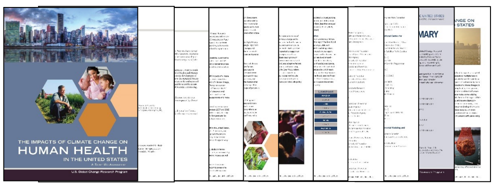
***Multi-stage***

***Context retrieved***: editing articles about museums and history. Featured content X Add a citation Current events Random article Automatic Manual Donate to Wikipedia Privacy policy About Wikipedia Disclaimers Contact Wikipedia Developers Cookie statement Reuse Wikipedia store URL, DOI, ISBN, PMC/PMID, QID. title, or citation Interaction ……Generate Help Click Insert Media to find a picture to add to your page. Type something to see suggestions. SecrelTunnel Talk Sandbox Preferences Beta Watchlist Contributions Log out User page Talk ……
***Question :*** What is the key to find a picture?
***Answer :*** insert media
***Real answer:*** Insert Media

***Single-stage***

***Context retrieved:*** Review your changes contributions To explore some more, click edit to make more changes. For instance, you might like to click the pencil button to see what Source editor looks like…… in editing articles about museums and history Featured content Add a link Done Current events Random article Donate to Wikipedia Privacy policy About Wikipeda Declamers Contact Wikipeda Developers Cookie stat Wikipedia External site Wikipedia store Museums Interaction Help Museum About Wikipedia Institution that holds artifacts and other obj ……
***Question :*** What is the key to find a picture?
***Answer :*** the magnifying glass icon
***Real answer:*** Insert Media

Figure 11: Case comparison of different retrieval strategies. The single-stage retrieval locates an error message.



***Multi-stage:***

***Context retrieved***: J. Balbus, J.L. Gamble, C.B. Beard, J.E. Bell, D. ……
PHOTO CREDITS port-on-the-impacts-of-climate-change-on-human-health-in cover and title page-Manhattan skyline: iStockPhoto.com/ the-united-states stockelements; Farmer: Masterfile/Corbis; Girl getting checkup: Rob Lewine/Tetra Images/Corbis 3. 2014: Climate Change Impacts in the United States: The Third Pg. vii-Elderly Navajo woman and her niece ……
Ch.9: Human health. Climate Change Impacts in the United States: The Third National Climate Assessment. Melillo, J.M., T..',
***Question :*** which country specified in this document?
***Answer :*** united states
***Real answer:*** United States

***Single-stage:***

***Context retrieved:*** Centers for Disease Control and Prevention Rupa Basu, California Office of Environmental Health Hazard Ross Bowling, Office of the Assistant Secretary for Administration Assessment Kathleen Danskin, Office of the Assistant Secretary for Preparedness Paul English, Public Health Institute, Oakland, CA and Response Kim Knowlton, Columbia University Mailman School of Public Health Stacey Degrasse ……
medium consensus Unlikely Medium Two kinds of language are used when describing the 1 in 3 Suggestive evidence …… Office of Air and Radiation Michelle Hawkins, National Oceanic and Atmospheric Am.
***Question :*** which country specified in this document?
***Answer :*** none
***Real answer:*** United States

Figure 12: Case comparison of different retrieval strategies. The single-stage retrieval could not locate the relevant information.