# Detect, Retrieve, Comprehend: A Flexible Framework for Zero-Shot Document-Level Question Answering

Tavish McDonald[1,*], Brian Tsan[2], Amar Saini[1], Juanita Ordonez[1], Luis Gutierrez[1], Phan Nguyen[1], Blake Mason[1] and Brenda Ng[1]

[1]*Lawrence Livermore National Laboratory, 7000 East Avenue Livermore, California 94550*
[2]*University of California Merced, 5200 Lake Rd, Merced, CA 95343*

**Abstract**
Researchers produce thousands of scholarly documents containing valuable technical knowledge. The community faces the laborious task of reading these documents to identify, extract, and synthesize information. To automate information gathering, document-level question answering (QA) offers a flexible framework where human-posed questions can be adapted to extract diverse knowledge. Finetuning QA systems requires access to labeled data (tuples of context, question and answer). However, data curation for document QA is uniquely challenging because the context (i.e., text passage containing evidence to answer the question) needs to be *retrieved* from potentially long, ill-formatted documents. Existing QA datasets sidestep this challenge by providing short, well-defined contexts that are unrealistic in real-world applications. We present a three-stage document QA approach: (1) text extraction from PDF; (2) evidence retrieval from extracted texts to form well-posed contexts; (3) QA to extract knowledge from contexts to return high-quality answers – extractive, abstractive, or Boolean. Using the QASPER dataset for evaluation, our *Detect-Retrieve-Comprehend* (DRC) system achieves a +7.19 improvement in Answer-$F_1$ over existing baselines due to superior context selection. Our results demonstrate that DRC holds tremendous promise as a flexible framework for practical scientific document QA.

**Keywords**
document understanding, information retrieval, question answering,

## 1. Introduction

Growth in new machine learning publications has exploded in recent years, with much of this activity occurring outside traditional publication venues. For example, arXiv hosts researchers' manuscripts detailing the latest progress and burgeoning initiatives. In 2021 alone, over 68,000 machine learning papers were submitted to arXiv. Since 2015, submissions to this category have increased yearly at an average rate of 52%. While it is admirable that the accelerated pace of AI research has produced many innovative works and manuscripts, the sheer amount of papers makes it prohibitively difficult to keep pace with the latest developments in the field. Increasingly, researchers turn to scientific search engines (e.g., Semantic Scholar and Zeta Alpha), powered by neural information retrieval, to find relevant literature. To date, scientific search engines [1, 2, 3] have focused on serving recommendations based on semantic similarity and lexical matching between a query phrase and a collection of document-derived contents, particularly titles and abstracts. Other efforts to elicit the details of scholarly papers have extracted quantified experimental results from structured tables [4] and generated detailed summaries from the hierarchical content of scientific documents [5].

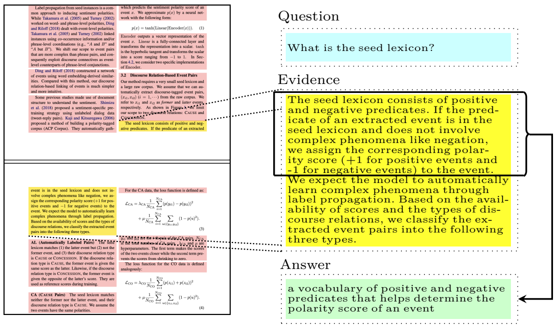


**Figure 1:** QASPER questions require PDF text extraction and evidence retrieval to generate an answer.

While these scientific search engines suffice for topic exploration, once a set of papers are identified as relevant, researchers would want to probe deeper for information to address specific questions conditioned on their prior domain knowledge (e.g., *What baselines is the neural relation extractor compared to?*). While one can gain a sense of the main findings of a paper by reading the abstract, the answers to these probing questions are frequently

*SDU'23: The Third AAAI Workshop on Scientific Document Understanding, February 14, 2023, Washington, D.C.*
*Corresponding author.
mcdonald53@llnl.gov (T. McDonald); btsan@ucmerced.edu (B. Tsan); saini5@llnl.gov (A. Saini); ordonez2@llnl.gov (J. Ordonez); gutierrez74@llnl.gov (L. Gutierrez); nguyen97@llnl.gov (P. Nguyen); mason35@llnl.gov (B. Mason); ng30@llnl.gov (B. Ng)

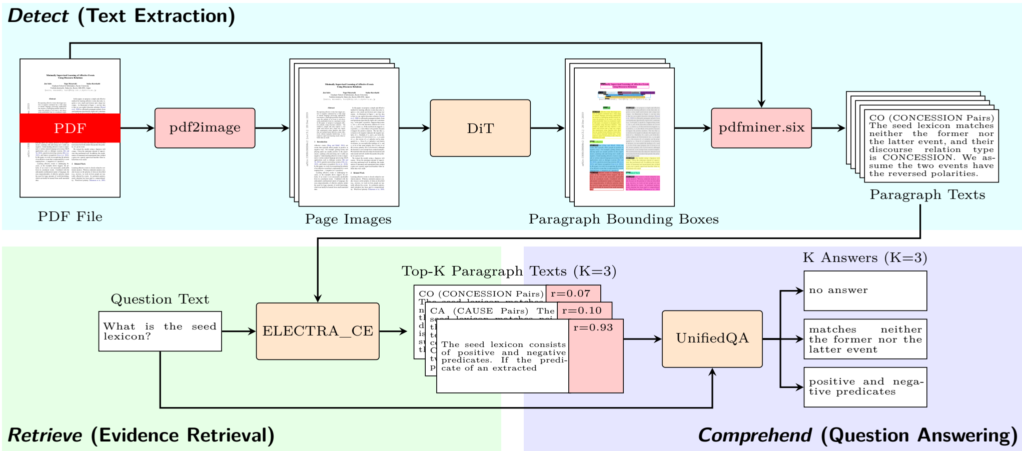


**Figure 2:** An instance of our modular end-to-end DRC system comprised of DiT + ELECTRA_CE + UnifiedQA.

found in the details of the methodology, experimental setup, and results sections. Furthermore, questions may require synthesis of document passages to produce an abstractive answer rather than simply extracting a contiguous span. Reading and manually cross-referencing the results of several papers is a labor-intensive approach to glean specific knowledge from scientific documents. Therefore, effective tools to help automate knowledge discovery are sorely needed.

A promising approach to extracting knowledge from scientific publications is document-level question answering (QA): using an open set of questions to comprehend figure captions, tables, and accompanying text [6]. Traditionally, the NLP community has focused on using *clean* texts as context to their QA systems. However, this is not representative of the vast majority of scholarly information found in structured documents. As QA garners interest from the computer vision community, DocVQA [7] and VisualMRC [8] have extended document QA to extracting evidence from single images, paving the way to extend contexts from text to visual sources.

A foundational challenge in building robust document QA systems is ensuring well-formed contexts, which entails accurate text extraction and requires adaptation to new document layouts. Nonetheless, even when text can be cleanly extracted, there still remains the crucial task of identifying question-relevant paragraphs for answer prediction.

Our contribution is a **general-purpose system for QA on full documents in their original PDF form**, that addresses the key challenges of scientific document QA: (1) accurate text extraction from unseen layouts, (2) evidence retrieval (i.e., context selection), and (3) robust QA.

## 2. Dataset

The Question Answering on Scientific Research Papers (QASPER) dataset consists of 1,585 NLP papers sourced from arXiv, and is accompanied by 5,049 questions from NLP readers and corresponding answers from NLP practitioners. Papers in QASPER are cited by their arXiv DOIs, which we used to fetch the original PDF documents as input to our system, as our work is focused on knowledge extraction at the PDF level.

QASPER contains 7,993 answers categorized by answer type: *Extractive* (4142), *Abstractive* (1931), *Yes/No* (1110), and *Unanswerable* (810). Using only the *Extractive*, *Abstractive* and *Yes/No* answers, we match our model prediction to the most similar answer when a question has more than one answer, and report our performance accordingly.

QASPER is ideal for evaluating our proposed framework because it provides: (1) paragraph text and table information to evaluate our layout-analysis model (in its ability to cleanly extract document regions); (2) evidence paragraphs to validate, and optionally finetune, our evidence retrieval model (in its ability to retrieve good context paragraphs); and (3) ground-truth answers to assess the accuracy of our QA model (in its ability to answer the question given the context).

## 3. Methodology

Document QA on raw PDFs is necessary towards automating knowledge extraction from scientific corpora and has remained an unaddressed problem. To address this, we propose a flexible information extraction tool to alleviate laboriously searching for answers grounded in evidence. Our system combines: (1) a robust text detector for visually rich documents, (2) explicit passage retrieval for evidence selection, and (3) multi-format answer prediction. We used pretrained open-source machine learning models that are effective in a zero-shot setting. We also finetuned these models to improve our system's end-to-end performance.

## 3.1. Problem Description

Our work addresses evidence retrieval at the PDF level. Thus, our document QA task is defined as: given a question and a PDF document, predict the answer to the question. We decompose this problem into three subtasks: text extraction (§ 3.2), evidence retrieval (§ 3.3), and QA (§ 3.4).

First, the PDF document, represented as a series of images, has its semantic regions identified and their corresponding text content extracted as passages. Second, the passages are ranked by their relevance to the question. Irrelevant passages are filtered out so only the most relevant passages are used as contexts for QA. Finally, given a context and question, the answer is predicted. The overall architecture is shown in Figure 2. These components correspond to the respective tasks of *Detect*, *Retrieve* and *Comprehend*, or DRC, which is also the name of our proposed system.

## 3.2. Detect

The first step of our pipeline is to extract text from PDF documents. Libraries such as *pdfminer.six* [9] and *TesseractOCR* [10] extract text from documents indiscriminately, including unwanted page numbers and footnotes, which would need to be filtered out before the extracted text can be used as context paragraphs. Thus, prior to text extraction, document layout analysis should be performed to detect targeted regions (from which text is to be extracted).

Document layout analysis models are trained to segment a document into its constituent components (e.g., paragraphs, figures, and tables). The Document Image Transformer (DiT) [11] is designed for layout analysis and text detection. DiT uses a masked image modeling objective to pretrain a Vision Transformer [12] without labels. It supports prediction of semantic region bounding boxes and segmentation masks. Predicted regions are then passed to OCR tools for text extraction.

In the *Detect* stage of DRC, the *pdf2image* library first converts each page of the document to images. For each image, DiT detects the bounding boxes for paragraphs. The text within each bounding box is then extracted using *pdfminer.six*. The extracted texts form passages which are candidates for question evidence.

## 3.3. Retrieve

Evidence retrieval identifies relevant passages by ranking them according to their similarity to the question. We considered several architectures.

**Lexical Retriever** BM25 [13] ranks questions and passages based on token-matching between sparse representations of the question and passage. Prior work has shown that BM25 is a strong baseline across many datasets [14, 15]. Given a question $q$ containing tokens $q_1, \ldots, q_T$ and a set of passages $\mathscr{P}$, the BM25 retrieval score $S$ between $q$ and passage $p \in \mathscr{P}$ is defined using TF-IDF token weights:

$$S_{q,p}^{BM25} = \sum_{i=1}^{T} \log\Big(\frac{|\mathscr{P}|}{N(q_i, \mathscr{P})}\Big) \frac{n(q_i, p)(k_1 + 1)}{k_1\big(1 - b + \frac{b|p|}{avpl}\big) + n(q_i, p)}$$

where $|\mathscr{P}|$ is the number of passages in the corpus; $|p|$ is the length of the passage; $N(q_i, \mathscr{P})$ is the number of passages with token $q_i$; $n(q_i, p)$ is the term frequency of $q_i$ in passage $p$; $avpl$ is the average passage length. $k_1$ and $b$ are constants.

**Dual Encoder** The Dense Passage Retriever (DPR) [16] learns via a contrastive training objective with in-batch negatives and hard negatives chosen by BM25. For question $q$ and a set of passages $\mathscr{P}$, DPR measures question-passage similarity with a dual-encoder architecture, where $f_q$ encodes the question $q$ and $f_p$ encodes the passage $p \in \mathscr{P}$ to the same latent space [17, 18]. The retrieval score $S$ is defined as the dot product of the two resulting embeddings:

$$S_{q,p;\Phi}^{DE} = f_q(q; \Phi_q) f_p(p; \Phi_p)$$

$\Phi = [\Phi_q, \Phi_p]$ denotes the retriever question and passage encoder parameters. We used the DPR *multi* variant, which has been trained on additional data, as Karpukhin et al. [16] has shown that the additional data improves retrieval generalizability.

**Cross-Encoder** Instead of embedding the question $q$ and passage $p$ separately, cross-encoders [19] compute a retrieval score $S$ where $f(q, p; \Phi)_{[\mathrm{CLS}]}$ encodes both question and passage using the CLS token representation of their concatenation:

$$S_{q,p;\Phi}^{CE} = \mathrm{softmax}\Big(f(q, p; \Phi)_{[\mathrm{CLS}]} W + b\Big)$$

where $W$ and $b$ are the weight and bias in the final layer that classifies whether $p$ is relevant to $q$. Many cross-encoders $f$ have since been proposed and a comparative analysis was performed [20], where the ELECTRA-base [21] cross-encoder (ELECTRA_CE) was declared as the

**Table 1**
Tuned hyperparameter values for the number of epochs, weight decay (WD), and batch size (BS) for finetuning. The learning rate for all trainable models is 2e-5.

| Type | Model | Epochs | WD | BS |
|---|---|---|---|---|
| Retriever | BM25 | – | – | – |
| | DPR-FT | 20 | 0 | 8 |
| | ELECTRA_CE-FT | 6 | 0.01 | 8 |
| Reader | UNIFIEDQA-FT | 10 | 0.01 | 10 |

**Table 2**
UNIFIEDQA Answer-$F_1$ scores using the top ranked context from extracted PDF regions.

| | | | Extractive | | Abstractive | | Boolean | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|
| Extractor | Retriever | QA | Val. | Test | Val. | Test | Val. | Test | Val. | Test |
| *TesseractOCR* | ELECTRA_CE | UNIFIEDQA | 34.47 | 34.77 | 21.05 | 21.99 | 74.69 | 73.06 | 33.82 | 34.75 |
| DIT + *pdfminer.six* | ELECTRA_CE | UNIFIEDQA | 33.91 | 34.65 | 21.80 | 22.18 | 74.97 | 78.73 | 34.05 | 35.28 |
| DIT + *pdfminer.six* | ELECTRA_CE | UNIFIEDQA-FT | 38.46 | 39.70 | 25.23 | 24.24 | **84.95** | **85.29** | 38.81 | 39.22 |
| DIT + *pdfminer.six* | ELECTRA_CE-FT | UNIFIEDQA | 33.70 | 35.11 | 22.89 | 22.74 | 77.28 | 79.90 | 34.63 | 35.88 |
| DIT + *pdfminer.six* | ELECTRA_CE-FT | UNIFIEDQA-FT | **39.18** | **41.79** | **26.29** | **25.49** | 84.02 | 81.82 | **39.46** | **40.16** |
| – | – | LED-base | 28.10 | 32.50 | 16.70 | 14.91 | 61.82 | 69.05 | 28.94 | 32.97 |
| – | – | LED-base-scaff | 23.37 | 29.59 | 15.49 | 14.95 | 66.36 | 67.14 | 26.37 | 31.59 |

best cross-encoder due to its stability and effectiveness across datasets. Thus, the ELECTRA_CE model (trained on MS MARCO [22]) was selected as our starting cross-encoder.

Once the passages have been ranked, the top-$K$ most relevant passages are used as contexts in the QA stage.

### 3.4. Comprehend

The final stage of DRC is comprehending a document's contents via *multi-format* question answering. These formats correspond to answer types, which can be extractive, abstractive, or Boolean. (An extractive answer is a span of text taken verbatim from an evidence passage. An abstractive answer is a generated span not quoted verbatim from the evidence. A Boolean answer is a binary prediction: yes or no.)

For comprehension, we use UNIFIEDQA [23], a generative question-answering model that has been pretrained on 20 datasets and can predict all answer formats with a single architecture. The answer type returned by UNIFIEDQA depends on the way the question is phrased. For each of the $K$ relevant passages (from the *Retrieve* stage), we pair the passage with the question as input to UNIFIEDQA to predict an answer. At the end, we have $K$ answers – one for each of the $K$ passages.

## 4. Experimental Setup

### 4.1. QASPER Baselines

Following Dasigi et al. [24], in our QASPER experiments, we use the Longformer-Encoder-Decoder (LED) [25] as the baseline model for evidence retrieval and QA. This model uses a modification of self-attention from the Transformer architecture [26] to encode longer sequences more efficiently. To jointly answer questions and decide whether a context is relevant in providing answer evidence, LED optimizes a multi-task objective. In addition to answer generation, LED adds a classification head (termed *evidence scaffold*) that operates over each paragraph to predict binary labels (evidence or non-evidence). Since we discarded unanswerable questions from QASPER, we retrain LED on the remaining questions and evaluate with and without evidence scaffolding. The retrained LED serves as a fairer competitor to UNIFIEDQA, which was not pretrained on unanswerable questions.

### 4.2. Text Extraction with Layout Analysis

We use DIT with *pdfminer.six* for selective text extraction. First, a pretrained DIT model predicts the bounding boxes of paragraphs on each page. Then, *pdfminer.six* extracts text within the bounding boxes. We denote this two-step procedure as DIT+*pdfminer.six*, and compare against *TesseractOCR*, which takes an image as input and returns the text found within the image, as well as *pdfminer.six*'s high-level extractor (*pdfminer.six**), which takes a PDF as

**Table 3**
Token extraction from paragraphs and tables within all documents from QASPER. $P$ denotes precision and $R$ denotes recall.

| Category | Method | $P$ | $R$ | $F_1$ |
|---|---|---|---|---|
| Paragraphs | DiT + *pdfminer.six* | **68.28** | 86.91 | **75.34** |
| | *pdfminer.six** | 48.95 | **90.31** | 62.54 |
| | *TesseractOCR* | 49.27 | 89.75 | 62.73 |
| Tables | DiT + *pdfminer.six* | **67.72** | 82.87 | **70.81** |
| | *pdfminer.six** | 6.88 | 92.71 | 12.38 |
| | *TesseractOCR* | 7.31 | **96.40** | 13.13 |

**Table 4**
Retrieval recall measured as the top percentages (top-$K\%$) of retrieved passages that contain the answer, averaged across all questions. The *Training* column makes explicit which retrievers are applied zero-shot or finetuned using the QASPER train split. For each pairing of data split and $K$, the best performing model is shown in **bold** and the second best is <u>underlined</u>.

| | | *K=1%* | | *K=5%* | | *K=10%* | | *K=20%* | |
|---|---|---|---|---|---|---|---|---|---|
| Retriever | Training | Val. | Test | Val. | Test | Val. | Test | Val. | Test |
| BM25 | – | 15.32 | 15.97 | 33.00 | 31.81 | 45.36 | 44.28 | 61.95 | 60.38 |
| DPR | – | 10.91 | 11.75 | 23.23 | 25.28 | 33.49 | 36.50 | 51.13 | 52.61 |
| ELECTRA_CE | – | <u>20.98</u> | <u>22.82</u> | <u>39.44</u> | <u>39.97</u> | <u>52.08</u> | <u>52.06</u> | <u>66.89</u> | <u>66.39</u> |
| DPR-FT | QASPER | **25.21** | **26.22** | 43.79 | 46.41 | 58.19 | 59.09 | 73.31 | 72.97 |
| ELECTRA_CE-FT | QASPER | 22.81 | 24.50 | **49.26** | **50.36** | **64.58** | **65.67** | **78.80** | **79.66** |

input and exploits PDF metadata to extract texts within the pages.

## 4.3. Retriever-QA Implementation Details

For BM25, we create an inverted index on QASPER validation and test sets using Pyserini [27] with default parameters ($k_1$=0.9, $b$=0.4). For DPR and ELECTRA_CE, we start with pretrained models from Hugging Face, then finetune them per hyperparameters shown in Table 1. In finetuning DPR and ELECTRA_CE, we sample batches containing a 1:4 ratio of positive to negative evidence passages.

For UnifiedQA, we use the unifiedqa-v2-t5-large-1363200 model from Hugging Face. We finetune it in a weakly supervised manner using evidence passages ranked by ELECTRA_CE but with the original questions and answers from QASPER. The choice to use retrieved passages (instead of the human-labeled evidence passages from QASPER) should make our system more robust to noisy context paragraphs. We show that a pretrained text extractor and evidence retriever can adapt UnifiedQA to the domain of QASPER papers without labeled evidence.

## 4.4. Evaluation Metrics

We evaluate DRC's text extraction, retrieval, and QA stages separately. For each stage, performance is measured by the $F_1$ score between the predicted outputs and the target labeled in QASPER. Adopting the same notation as Dasigi et al. [24], we name the $F_1$ scores for our evidence retrieval and QA as Evidence-$F_1$ and Answer-$F_1$, respectively. For text extraction, in addition to its $F_1$ score, we also evaluate its precision and recall.

Since each question in QASPER is labeled with its answer(s) and accompanying evidence, it is possible to evaluate both our QA and evidence retrieval stages using this single dataset. For QA, Answer-$F_1$ is calculated between the tokens in the predicted answer and the tokens in the target answer. For our evidence retrieval stage, which ranks passages by their relevance to a given question, Evidence-$F_1$ is calculated between a fixed percentage of the top ranking passages and the set of passages labeled as evidence in QASPER.

QASPER also contains the plain text for each of its documents, organized so that text from paragraphs and tables are separated. We use this plain text to evaluate the efficacy of our text extraction to extract only the primary content of PDF document. The precision, recall, and $F_1$ score for text extraction are calculated between the tokens in a document's extracted text and its tokens in QASPER's plain text version. For all of our experiments, tokenization is performed at the word level, using our pretrained UnifiedQA model's tokenizer.

**Table 5**
UnifiedQA Answer-$F_1$ scores on UnifiedQA using the top ranked context from selected retrievers compared to the LED baselines. For all data splits, the best performing model is shown in **bold** and the best zero-shot model is underlined.

| Retriever | QA | Extractive Val. | Extractive Test | Abstractive Val. | Abstractive Test | Boolean Val. | Boolean Test | Overall Val. | Overall Test |
|---|---|---|---|---|---|---|---|---|---|
| BM25 | UnifiedQA | 29.24 | 30.04 | 23.12 | <u>24.77</u> | <u>77.40</u> | 78.99 | 33.23 | 36.85 |
| DPR | UnifiedQA | 28.28 | 28.94 | 23.60 | 21.85 | 75.96 | 79.38 | 32.58 | 35.59 |
| ELECTRA_CE | UnifiedQA | <u>32.72</u> | <u>34.46</u> | <u>24.34</u> | 23.40 | 77.30 | <u>80.63</u> | <u>35.73</u> | <u>39.51</u> |
| DPR-ft | UnifiedQA | **33.37** | **34.64** | **24.97** | 24.53 | **78.66** | 77.16 | **36.46** | 39.31 |
| ELECTRA_CE-ft | UnifiedQA | 33.18 | 34.21 | 24.56 | **25.78** | 78.20 | **81.45** | 36.18 | **40.03** |
| – | LED-base | 28.10 | 32.50 | 16.70 | 14.91 | 61.82 | 69.05 | 28.94 | 32.97 |
| – | LED-base-scaff | 23.37 | 29.59 | 15.49 | 14.95 | 66.36 | 67.14 | 26.37 | 31.59 |

**Table 6**
Comparison between ELECTRA cross-encoders against LED baselines in terms of Evidence-$F_1$.

| Model | Evidence-$F_1$ Val. | Evidence-$F_1$ Test |
|---|---|---|
| ELECTRA_CE | **31.75** | 36.37 |
| ELECTRA_CE-ft | 31.58 | 36.12 |
| LED-base | 23.94 | 29.85 |
| LED-base-InfoNCE | 24.90 | 30.60 |
| LED-large | 31.25 | **39.37** |

# 5. Results

We demonstrate DRC's effectiveness on document QA by measuring its end-to-end performance. We also evaluate its constituent components on text detection, evidence retrieval, and QA tasks against existing QASPER baselines. For evidence retrieval, we study the benefits of having a separate retrieval process, in contrast to the evidence selection scaffold for LED. Furthermore, we explore DRC's performance in both zero-shot and finetuned settings, to assess its performance under varying degrees of access to labeled data.

## 5.1. End-to-End QA System

Table 2 shows DRC's performance in terms of Answer-$F_1$. In these experiments, we extract text from documents using either *TesseractOCR* or DiT and rank passages using ELECTRA_CE. We then pass highly-ranked passages to UnifiedQA for answer prediction. First, we study the influence of the text detection model on Answer-$F_1$ performance by comparing *TesseractOCR* to DiT. While we observe that using DiT reports higher Answer-$F_1$ than *TesseractOCR* across all answer types, the difference is negligible.

Next, we examine DRC in the zero-shot setting where ELECTRA_CE and UnifiedQA models are not finetuned. DRC achieves an overall +2.31 improvement in Answer-$F_1$ over LED-base without scaffolding on QASPER's test split. To improve upon the fully zero-shot approach, we apply weak supervision to finetune UnifiedQA: we sample extracted passages according to their retrieval scores from the pretrained ELECTRA_CE model, assuming that higher ranked passages are correlated with selection probability for answer prediction. Thus, we are able to finetune UnifiedQA without access to human-labeled contexts, since labeled question-answer pairs are generally unavailable for large technical corpora. This finetuning approach yields a +6.25 improvement to LED-base in overall Answer-$F_1$ on the test dataset.

To analyze Answer-$F_1$ performance when ground-truth question-passage pairs are available, we consider an ELECTRA_CE retriever finetuned on QASPER's training set. We then finetune UnifiedQA through weak supervision using the now improved retriever. DRC with a finetuned ELECTRA_CE shows modest gains over the zero-shot system but still lesser performance compared to the pretrained ELECTRA_CE with a weakly-supervised UnifiedQA. This suggests that downstream QA performance is better improved by adapting to the target domain QASPER documents, than by receiving more relevant passages.

Combining a finetuned ELECTRA_CE retriever with a weakly supervised UnifiedQA model shows the greatest improvements over LED-base without scaffolding, +7.19 in Answer-$F_1$ on the and test dataset for all answer types. We observe that using finetuned ELECTRA_CE for weak supervision shows worse performance on Boolean questions than using the pretrained ELECTRA_CE to weakly supervise UnifiedQA. This discrepancy is likely due to the small proportion of Boolean samples in the validation and test datasets compared to other formats, 13% and 15% respectively.

Across all experiments, DRC demonstrates superior performance to LED while solving a more difficult task: DRC starts from PDFs while LED starts from clean texts. DRC bridges an essential gap in real-world applications for scientific knowledge extraction because PDFs are directly processed as input. In the following discussion, we validate our individual system components.

## 5.2. Text Detection

We evaluate three different methods for text extraction from PDF files: (1) DiT+*pdfminer.six*, (2) *pdfminer.six**, and (3) *TesseractOCR*. Table 3 reports the average precision, recall, and $F_1$ between the extracted tokens and those in the ground-truth text.

For paragraph extraction, DiT+*pdfminer.six* has better precision than *pdfminer.six** (+19.33) and *TesseractOCR* (+19.01). We attribute this improvement to extracting fewer unwanted artifacts (e.g., page numbers, headers, footers, and footnotes). For text within tables, only DiT+*pdfminer.six* is effective off-the-shelf. *pdfminer.six** and *TesseractOCR* do not disambiguate between text in and outside of tables. *pdfminer.six** and *TesseractOCR* would suffice if text is contained only in tables, or only in paragraphs, but not a mixture of the two because the text from tables and paragraphs will be interspersed.

## 5.3. Evidence Passage Retrieval

We compare DPR, ELECTRA_CE, and BM25 by their ability to rank passages by relevance to questions. Table 4 shows the recall of evidence passages within various percentages of the top ranked passages, averaged over all questions in QASPER, for the retrievers in both zero-shot and finetuned settings. As questions are posed for a specific document, our retrievers consider a variable number of passages per question because documents vary in length. Since top-$K$ penalizes longer documents when $K$ is small, we measure recall using top-$K$% for $K \in \{1, 5, 10, 20\}$.

In the zero-shot setting, BM25 outperforms DPR by an average of +15.58 gain in recall on the test data. These results support findings from Sciavolino et al. [15], who reported that DPR trained on Natural Questions [28] underperformed BM25 when faced with the new question patterns and entities found in their EntityQuestions dataset. Thus, DPR requires finetuning and is less generalizable than BM25, which has no trainable parameters.

ELECTRA_CE, with average recall gains of +7.2 and +13.78 over BM25 and DPR, respectively, is the clear winner. We hypothesize that ELECTRA_CE's success is due to the explicit interaction between every token of the question and passage through its cross-attention mechanism, offering a more expressive similarity function than DPR's inner product between question and passage or BM25's weighted term matching.

To analyze how retrievers perform with ground-truth question-passage pairs, we also evaluate passage retrieval with DPR and ELECTRA cross-encoders finetuned on QASPER. Here, ELECTRA_CE outperforms DPR on the test data for $K$ = 5%, 10% and 20% by an average recall of +3.95, +6.58, and +6.69, respectively. Notably, DPR has higher recall for $K$=1%. We conjecture that this may be due to DPR's contrastive objective utilizing hard negative sampling, but further analysis on the relationship between training objective and ranking is needed.

### 5.3.1. Comparison to Evidence Selection Scaffold

To compare against LED's evidence scaffold, we now treat ELECTRA_CE as a binary classifier. Akin to LED's evidence scaffold, we use the [CLS] representation of the question-passage pair as input to a single layer neural network to estimate the probability that the passage is relevant as evidence to the question and use a classification threshold of 0.5 [19]. Table 6 illustrates the evidence classification performance of zero-shot and finetuned ELECTRA_CE models against LED variations. Evidence-$F_1$ scores are computed using the extracted passages classified as evidence with respect to the ground-truth set labeled in QASPER. We observe that the difference between the zero-shot and finetuned ELECTRA_CE models is negligible. On the test split, zero-shot ELECTRA_CE shows a notable Evidence-$F_1$ improvement over LED-base augmented with InfoNCE loss [29], but is outperformed by LED-large. This agrees with findings from Dasigi et al. [24] that LED-large generally outperforms LED-base for retrieval but not QA. Thus, we consider only LED-base in our subsequent experiments with downstream QA.

## 5.4. Question Answering

Here, we focus on the effect of using retrieval mechanisms on QASPER's plain text passages for answer prediction. We report Answer-$F_1$ scores for extractive, abstractive, and Boolean answer types. We also finetune DPR and ELECTRA_CE models on QASPER's train split and compare against LED variations.

Table 5 shows UnifiedQA's Answer-$F_1$ using the highest-ranked passage from each retriever. We observe that UnifiedQA (first 5 rows) generally yields higher Answer-$F_1$ scores across answer types, datasets, and retrievers than LED baselines (last 2 rows). The exception is extractive answers from the test set where BM25 reports a lower Answer-$F_1$ than LED-base. Similarly, a zero-shot DPR retriever performs worse than both LED models. Among zero-shot retrievers, ELECTRA_CE yields the best performance on the overall test set with a +7.92 and +6.54 Answer-$F_1$ increase over LED with and without

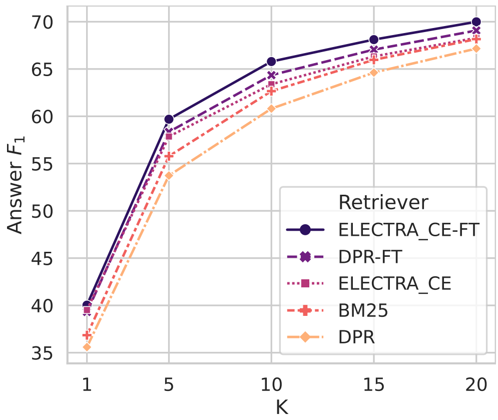


**Figure 3:** UnifiedQA overall Answer-$F_1$ performance on the QASPER test data split using the top-$K$ ranked contexts from selected retrievers. Retrievers with "-FT" denote those finetuned on QASPER's train split.

scaffolding. While DPR benefits the most from finetuning, the finetuned ELECTRA_CE reports the highest test performance across all answer types.

To verify that a weakly-supervised UnifiedQA model improves end-to-end answer prediction using weakly-labeled evidence, we measure the effect of retrieval beyond the top-ranked passage. We report UnifiedQA's overall Answer-$F_1$ on the test data using the best answer predicted from the top-$K$ ranked passages. Figure 3 compares performance using zero-shot and finetuned retrievers. Zero-shot ELECTRA_CE consistently dominates BM25 and DPR without finetuning. When finetuned, ELECTRA_CE reports the best performance for all choices of $K$. This confirms our hypothesis that we can adapt UnifiedQA to achieve higher quality answers, via weak supervision using ranking signals from the retriever.

## 6. Conclusion

We introduced DRC, an end-to-end QA system for automating manual knowledge extraction from scientific PDF documents. We showed that DRC greatly improves over existing baselines, which act on clean texts and sidestep the challenge of PDF-to-text extraction. Through extensive experiments, we evaluate our pipeline components in both zero-shot and finetuned settings. In practice, datasets as comprehensive as QASPER are few and may not be feasible for niche domains. In such cases, a fully zero-shot pipeline is mandatory for document QA, and DRC can be weakly supervised to adapt to specific domains. Our DRC sets a new benchmark for QASPER and serves as a proof of concept for an end-to-end document QA system, from PDF to answer. Key takeaways from our experiments include:

1. DiT demonstrates superior text extraction performance to *pdfminer.six* and *TesseractOCR*.
2. Zero-shot ELECTRA_CE offers the best retrieval performance for all top-$K$% (where $K \in \{1, 5, 10, 20\}$).
3. DRC adapts to new domains through weakly-supervised training on evidence passages leading to substantially improved answer prediction over LED baselines.

In this work, we have only scratched the surface with text. Future QA systems should be able to process scientific documents with diverse layouts and visually rich content. In order to fully automate information extraction beyond text, we must augment our system to identify and understand visual elements (e.g., figures) by incorporating visual question answering (VQA) and multimodal representations. Additionally, QASPER consists of well-formatted research papers in digitally-generated PDF documents. Further experiments are required to evaluate DRC's performance under domain shifts with respect to document format and text cleanliness (e.g., OCR noise).

Deployed document QA systems must operate on documents of varying lengths. We demonstrated retrieval and evidence selection on relatively short research papers, whose lengths are unrepresentative of manual processing tasks that require comprehending entire textbooks and technical manuals. As the number of passages grows, the input sequence limit will be exceeded, making answer prediction and evidence selection by efficient transformer architectures challenging. In addition, ranking via cross-encoder may be computationally expensive due to the costly cross-attention operation between the question and each passage. Thus, QA on lengthy documents may require a dual-encoder retriever, to store precomputed passage embeddings using FAISS [30] to maximize efficiency.


## Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.


## References


[1] M. Fadaee, O. Gureenkova, F. Rejon Barrera, C. Schnober, W. Weerkamp, J. Zavrel, A new neural search and insights platform for navigating and organizing AI research, in: Proceedings of the First Workshop on Scholarly Document Processing, Association for Computational Linguistics, Online,

2020, pp. 207–213. URL: https://aclanthology.org/2020.sdp-1.23. doi:10.18653/v1/2020.sdp-1.23.

[2] T. Zhao, K. Lee, Talk to papers: Bringing neural question answering to academic search, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 30–36. URL: https://aclanthology.org/2020.acl-demos.5. doi:10.18653/v1/2020.acl-demos.5.

[3] M. Parisot, J. Zavrel, Multi-objective representation learning for scientific document retrieval, in: Proceedings of the Third Workshop on Scholarly Document Processing, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022. URL: https://aclanthology.org/2022.sdp-1.9.

[4] M. Kardas, P. Czapla, P. Stenetorp, S. Ruder, S. Riedel, R. Taylor, R. Stojnic, AxCell: Automatic extraction of results from machine learning papers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 8580–8594. URL: https://aclanthology.org/2020.emnlp-main.692. doi:10.18653/v1/2020.emnlp-main.692.

[5] S. Sotudeh, A. Cohan, N. Goharian, On generating extended summaries of long documents, arXiv preprint arXiv:2012.14136 (2020).

[6] L. u. Borchmann, M. Pietruszka, T. Stanislawek, D. Jurkiewicz, M. Turski, K. Szyndler, F. Graliński, Due: End-to-end document understanding benchmark, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/069059b7ef840f0c74a814ec9237b6ec-Paper-round2.pdf.

[7] M. Mathew, D. Karatzas, C. Jawahar, Docvqa: A dataset for vqa on document images, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2200–2209.

[8] R. Tanaka, K. Nishida, S. Yoshida, Visualmrc: Machine reading comprehension on document images, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 13878–13888. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17635.

[9] Y. Shinyama, P. Guglielmetti, Pdfminer.six, https://github.com/pdfminer/pdfminer.six, 2020.

[10] R. Smith, An overview of the tesseract ocr engine, in: Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, ICDAR '07, IEEE Computer Society, USA, 2007, p. 629–633.

[11] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, F. Wei, Dit: Self-supervised pre-training for document image transformer, arXiv preprint arXiv:2203.02378 (2022).

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[13] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[14] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/65b9eea6e1cc6bb9f0cd2a47751a186f-Paper-round2.pdf.

[15] C. Sciavolino, Z. Zhong, J. Lee, D. Chen, Simple entity-centric questions challenge dense retrievers, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6138–6148. URL: https://aclanthology.org/2021.emnlp-main.496. doi:10.18653/v1/2021.emnlp-main.496.

[16] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.

[17] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a "siamese" time delay neural network, in: J. Cowan, G. Tesauro, J. Alspector (Eds.), Advances in Neural Information Processing Systems, volume 6, Morgan-Kaufmann, 1993. URL: https://proceedings.neurips.cc/paper/1993/file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf.

[18] D. Singh, S. Reddy, W. Hamilton, C. Dyer, D. Yogatama, End-to-end training of multi-document reader and retriever for open-domain question answering, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Informa-

tion Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 25968–25981. URL: https://proceedings.neurips.cc/paper/2021/file/da3fde159d754a2555eaa198d2d105b2-Paper.pdf.
[19] R. Nogueira, K. Cho, Passage re-ranking with bert, ArXiv abs/1901.04085 (2019).
[20] X. Zhang, A. Yates, J. Lin, Comparing score aggregation approaches for document retrieval with pretrained transformers, in: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2021, p. 150–163. URL: https://doi.org/10.1007/978-3-030-72240-1_11. doi:10.1007/978-3-030-72240-1_11.
[21] K. Clark, M. Luong, Q. V. Le, C. D. Manning, 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, 2020. URL: https://openreview.net/forum?id=r1xMH1BtvB.
[22] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, Ms marco: A human generated machine reading comprehension dataset, arXiv preprint arXiv:1611.09268 (2016).
[23] D. Khashabi, Y. Kordi, H. Hajishirzi, Unifiedqa-v2: Stronger generalization via broader cross-format training, arXiv preprint arXiv:2202.12359 (2022). https://huggingface.co/allenai/unifiedqa-v2-t5-large-1363200.
[24] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N. A. Smith, M. Gardner, A dataset of information-seeking questions and answers anchored in research papers, in: NAACL, 2021.
[25] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).
[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[27] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2356–2362. URL: https://doi.org/10.1145/3404835.3463238. doi:10.1145/3404835.3463238.
[28] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural questions: a benchmark for question answering research, Transactions of the Association of Computational Linguistics (2019).
[29] A. Caciularu, I. Dagan, J. Goldberger, A. Cohan, Utilizing evidence spans via sequence-level contrastive learning for long-context question answering, arXiv preprint arXiv:2112.08777 (2021).
[30] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data 7 (2019) 535–547.