

DocSeeker: Structured Visual Reasoning with Evidence Grounding for Long Document Understanding

Supplementary Material

A. Details of Data Distillation

A.1. Overview of the Distillation Pipeline

To address the issue of supervision scarcity in existing multi-page document VQA datasets, which typically provide only final short answers and evidence page indices without intermediate reasoning steps, we designed a rigorous data distillation pipeline. As illustrated in Figure A1, this pipeline transforms raw samples into high-quality structured ALR CoT data. The overall process consists of three key phases:

Data Filtering and Context Construction. We strictly select long-document samples from MP-DocVQA and DUDE, applying a pre-filtering step to discard documents that are overly short. Subsequently, we construct a "Minimal Context" for each sample, composed exclusively of the evidence pages, corresponding Page IDs, and the question. This strategy significantly reduces distillation costs while substantially enhancing the quality of generated data. Quantitative analysis on a subset of 1,000 long documents (avg. 17.6 pages) demonstrates that this approach improves the distillation success rate from 20.4% (full-context input) to 67.3%.

ALR CoT Data Distillation. We employ Gemini-2.5-Flash [39] as the teacher model due to its strong multi-modal reasoning capabilities. By utilizing a specialized prompt, we instruct the teacher to act as an expert annotator. The teacher generates a structured response that explicitly includes Question Analysis, Evidence Localization (citing specific Page IDs), and a Reasoning Process, strictly following our proposed ALR paradigm.

Secondary Verification. To guarantee the correctness of the synthesized reasoning paths, we implement a robust two-stage verification mechanism. First, we apply an automated Exact Match (EM) filter to check if the generated final answer and evidence page IDs align perfectly with the ground truth. For samples that fail the strict EM check (e.g., due to paraphrasing), we employ GPT-4o as a semantic judge to validate the correctness of the answer. Only samples passing this verification are retained for the SFT stage.

A.2. Prompt Templates

This section outlines the specific prompt templates integral to our data distillation pipeline. Figure A2 presents the ALR CoT Data Distillation Prompt, designed to extract high-quality structured reasoning from the teacher model.

Furthermore, we provide the Vanilla CoT Data Distillation Prompt in Figure A3 for ablation studies; unlike the ALR approach, this prompt encourages the model to generate free-form reasoning chains without enforcing the structured constraints of question analysis or explicit evidence localization. Finally, Figure A4 illustrates the LLM-based Judge Prompt, employed during secondary verification to salvage factually correct samples initially discarded due to formatting mismatches.

A.3. Failure Analysis

We visualize representative examples of samples that failed the initial automated Exact Match (EM) screening in Figure A5, which fall into two distinct categories:

1) **Genuine Reasoning Failures.** These cases contain substantive quality defects. Despite attempting to follow the ALR paradigm, the model commits logical errors during the reasoning process, resulting in factually incorrect answers. These constitute actual noise and must be strictly discarded.

2) **Correct Reasoning with Formatting Mismatches.** These samples exhibit correct analysis, localization, and reasoning process. The derived answers are semantically correct but fail the EM check solely due to discrepancies in formatting, e.g. 17th February 1916 vs 1916-02-17. Such samples also represent high-quality supervision signals that should be preserved to maximize data utilization.

These observations underscore the necessity of the Secondary Verification mechanism introduced in Section A.1, which is essential for accurately distinguishing valid semantic matches from genuine reasoning errors.

B. Additional Ablation Study Results

To delve deeper into the impact of distinct reasoning paradigms and data strategies on Generalization Capabilities, we supplemented the In-domain (DUDE [2]) evaluation in Table A1. SFT on Raw Short-Answer data yields In-domain gains comparable to other paradigms, yet improvement on OOD tasks remains negligible. This suggests the model primarily engages in Rote Memorization rather than mastering the general reasoning skills required for unseen long documents. In contrast, increasing the scale of ALR CoT data not only progressively enhances In-domain performance but, critically, achieves a synchronous leap in OOD tasks. This demonstrates that the structured "Analysis-Localization-Reasoning" paradigm instills transferable reasoning capabilities, enabling the model to transcend specific data distributions and actively adapt to un-

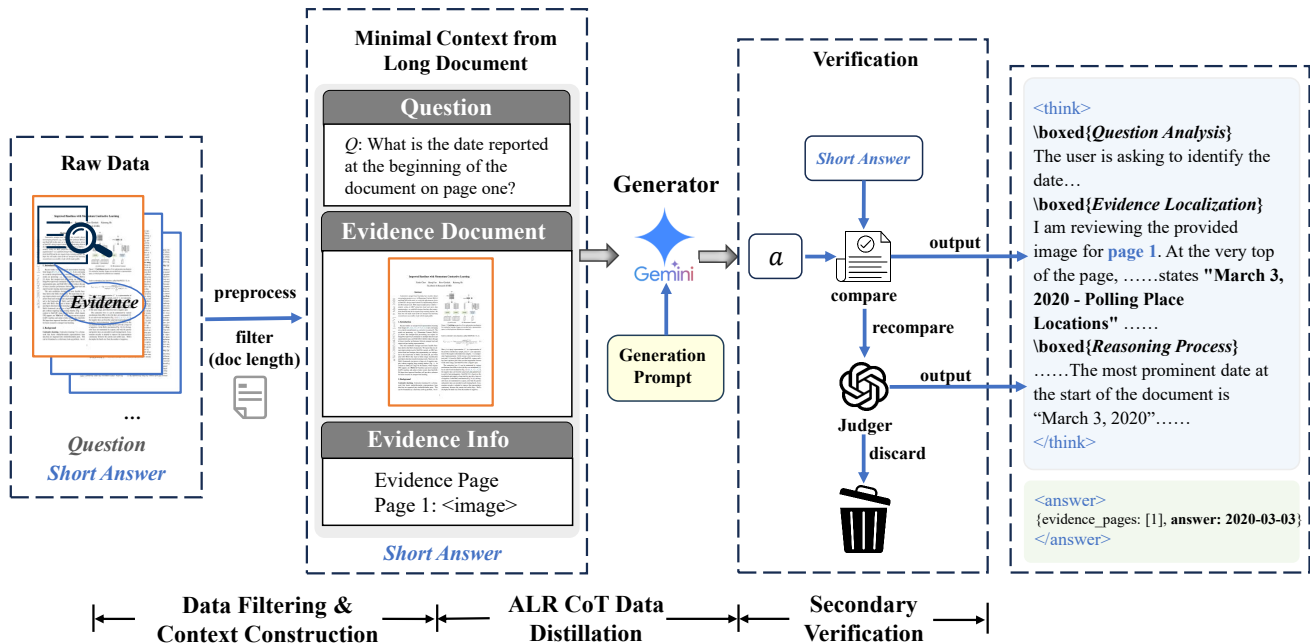


Figure A1. Data Distillation Pipeline.

Table A1. Ablation study on data types and training data size. The first group investigates different CoT data configurations, while the second explores the effect of varying data size on performance.

| Data Config | Size | MMLongbench. | | DUDE |
|-----------------------|------|--------------|-------------|-------------|
| | | Acc. | F1 | ANLS |
| Baseline | - | 25.4 | 20.8 | 35.2 |
| Raw short-answer data | 6.3k | 27.4 | 27.6 | 48.8 |
| Vanilla CoT data | 6.3k | 31.3 | 32.4 | 48.7 |
| ALR CoT data | 6.3k | 33.8 | 33.9 | 48.9 |
| -w/o Page id | 6.3k | 30.4 | 31.1 | 47.5 |
| 20% of ALR CoT data | 2.8k | 32.7 | 31.7 | 46.5 |
| 40% of ALR CoT data | 5.6k | 35.8 | 33.8 | 52.3 |
| 60% of ALR CoT data | 8.4k | 36.5 | 34.9 | 54.5 |
| 80% of ALR CoT data | 11k | 38.2 | 36.5 | 55.9 |
| All ALR CoT data | 13k | 38.6 | 36.9 | 56.5 |

familiar long-document structures, thereby achieving true generalization.

C. Efficiency and Accuracy Trade-offs

Although the proposed ALR paradigm enhances the model’s capability in long-document understanding, it inevitably introduces modest additional computational latency. To investigate the trade-off between inference latency and accuracy, we conducted experiments on MMLongBench-Doc, as shown in Table A2. Although the ALR paradigm nearly doubles output tokens, the end-to-

end latency increases only modestly from 19s to 25s. This is because the primary computational bottleneck in long-document processing is the visual pre-fill stage rather than the text decoding stage. Ultimately, DocSeeker trades a 31.6% latency increase for a 14.7% absolute accuracy gain and enhanced interpretability.

D. Stage-wise Error Breakdown

To systematically decompose the errors encountered in long-document understanding, we perform a stage-wise error breakdown on the samples from MMLongBench-Doc.

ALR CoT Data Distillation Prompt

Task Definition:

You are an AI assistant specialized in constructing high-quality question-answering datasets. Your core task is to generate a detailed reasoning process and a complete, natural answer based on the given question, the corresponding document evidence page, and the page number.

Important Rule:

All reasoning and the final answer must be strictly derived from the provided *Document Page Content*, without using external knowledge or unsupported guesses.

Output Requirements:

Your output must strictly contain two sections, “<think>” and “<answer>” following the exact format below.

Expected Output Format:

Reasoning Process (“<think>” section)

This section must use “\boxed{ }” to wrap the following three subheadings to clearly present your reasoning steps:

<think>

\boxed{Question Analysis}

Analyze the intent of the question in depth. Clearly identify what entity, concept, quantity, or relationship the question is asking about. Break complex questions down if needed.

\boxed{Evidence Localization}

Identify which page contains the evidence, then locate the key content within the provided Document Page Content.

\boxed{Reasoning Process}

Derive the answer step by step using only localized evidence. Explain how the conclusion is formed.

</think>

Final Answer (“<answer>” section)

This section must be a single JSON object containing:

<answer>

evidence_pages: A JSON list including the page numbers.

answer: A concise string giving the direct answer without introductory phrases or full sentence structures.

</answer>

Example:

Question: mention the texts in brackets with bold fonts in paragraph 2 in page 5?

Document Page Number: [5]

<think>

\boxed{Question Analysis}

The user asks for bracketed text in paragraph 2 on page 5 that is also bolded.

\boxed{Evidence Localization}

Page 5 contains bracketed text but no bolded text.

\boxed{Reasoning Process}

No text meets all required conditions; therefore the question cannot be answered.

</think>

<answer>

{"evidence_pages": [5], "answer": "Not answerable"}

</answer>

Figure A2. ALR CoT Data Distillation Prompt Template.

Vanilla CoT Data Distillation Prompt

Task Definition:

You are an AI assistant specialized in constructing high-quality question-answering datasets. Your core task is to generate a detailed reasoning process and a complete, natural answer based on the given question, the corresponding document evidence page, and the page number.

Requirements:

- All your reasoning and final answer must be strictly derived from the provided “Document Page Content”.
- You are not allowed to use any external knowledge or make unsupported guesses.
- Your output must strictly follow the format below with two required sections: “<think>” and “<answer>”.

Expected Output:

<think>

Let’s think step by step.

...

</think>

<answer>

A string that provides a direct and concise answer to the question without any introductory phrases or full sentence structures.

</answer>

Figure A3. Vanilla CoT Data Distillation Prompt Template.

Table A2. Analysis of inference latency and accuracy trade-off on MMLongBench-Doc

| | Token | Latency | Acc |
|-----------|-------|---------|------|
| Baseline | 202 | 19s | 25.4 |
| DocSeeker | 401 | 25s | 40.1 |

Table A3. Stage-wise error breakdown on MMLongBench-Doc.

| | Acc \geq 0.5 | Acc= 0 | Total |
|--------------|----------------|--------|-------|
| Recall= 1 | 300 | 288 | 588 |
| Recall< 1 | 121 | 373 | 494 |
| Total | 421 | 661 | 1082 |

As detailed in Table A3, we categorize these samples based on the correlation between evidence recall (Recall) and answer accuracy (Acc), which reveals three primary failure modes:

(i) **Reasoning failures.** The model successfully retrieves all required evidence pages ($Recall = 1$) but fails to provide the correct final deduction

(ii) **Localization failures.** The model fails to localize the evidence pages, leading to a lack of sufficient context for the model to perform downstream reasoning.

(iii) **Exceptions.** The model provides correct answers

despite incomplete localization ($Recall < 1$) by leveraging its internal global context awareness.

E. Details of Compared Methods

Table A4 details the specific evaluation configurations for the compared models across five benchmarks. For all baselines, we strictly adhered to the officially recommended or optimal settings to ensure a fair comparison.

F. Case Studies

In this section, we present qualitative visualizations of representative examples to demonstrate the superior capabilities of DocSeeker in handling complex long-document understanding tasks.

F.1. Comparative Analysis of Reasoning Paradigms

Figure A6 presents a qualitative comparison of DocSeeker against alternative reasoning paradigms. Experimental results indicate that the traditional Baseline model, constrained by limited long-context processing capabilities, exhibits complete localization failure, whereas the Short-Answer model yields “black-box” predictions devoid of logical support. Vanilla CoT successfully localizes evidence, it suffers from “reasoning drift” during the unstruc-

Judge Prompt

Task Definition:

You are an expert AI assistant for data validation and correction. Your core task is to compare a model-generated response with a ground-truth answer and determine the final, most accurate output based on a set of rules.

Inputs:

You will be given three pieces of information: *question*, *response*, and *answer*.

Requirements:

Your output should be a single, clean string representing the corrected answer, or the word 'Error' if applicable.

Rules:

1. **Formatting Mismatch:** If the response and answer contain the same core information but differ only in formatting (e.g., date separators like / vs. -, number separators like , vs. ., or case sensitivity), you must output the version from the ground-truth answer.
2. **Overly Detailed Response:** If the response is factually correct and contains all the information from the answer, but includes additional conversational phrases or details (e.g., "The answer is X" vs. "X"), you should conservatively output the more concise ground-truth answer.
3. **Overly Simplified Response:** If the response is correct but misses key details that are present in the answer, you must output the more complete ground-truth answer.
4. **Correct Response, Flawed Answer:** If you determine that the response is factually or grammatically more correct than the ground-truth answer (e.g., the answer has a clear typo or factual error that the response corrects), you must output the model-generated response to ensure data quality.
5. **Factually Incorrect/Contradictory:** If the response and answer are factually contradictory and irreconcilable (e.g., different names, numbers, dates), or if one provides a factual answer while the other states it's not answerable, you must output the exact string `Error`.

Output Format:

```
output :  
<answer>
```

Please read the following example and complete the output in the specified format based on the given information.

Example:

Question:

WHAT IS THE APPROVED DATE FOR RELEASE?

Response:

2001/07/27

Answer:

2001-07-27

Output:

2001-07-27

Figure A4. LLM-based Judge Prompt Template in Secondary Verification.

tured generation process, resulting in deviations within the logical chain. In contrast, DocSeeker leverages the structured ALR paradigm to strictly decouple evidence acquisition from logical deduction. This constraint mechanism

effectively mitigates reasoning instability and ensures the execution of rigorous inference steps based on precise evidence, thereby demonstrating the significant robustness of our paradigm in complex long-document understanding.

Failure Cases

Question

How many Adjudicatory Cases are there in the document?

Ground Truth

Answer: 12 Evidence_pages: [3, 4, 5] Evidence_sources: ['Table']

<think>

Question Analysis

The user is asking for the total count of "Adjudicatory Cases" present in the document. This requires identifying the section dedicated to adjudicatory cases and then counting each individual case listed within that section.

Evidence Localization

I will examine the provided document pages (3, 4, and 5). On page 3, there is a clear heading "ADJUDICATORY CASES". Under this heading, cases are listed, each identified by a "PCB" number. I will count each distinct entry under this section across all provided pages.

Reasoning Process

1. Locate the section: The "ADJUDICATORY CASES" section begins on page 3.
2. Count cases on Page 3:- PCB 21-3, PCB 21-5 (One combined entry)- PCB 21-29, -PCB 21-47 (page 3: 3 cases)
3. Count cases on Page 4: - PCB 21-110 - PCB 21-114 - PCB 22-11... **Flawed reasoning**
5. Sum the counts: 3 (page 3) + 7 (page 4) + 1 (page 5) = 11. Therefore, there are 11 adjudicatory cases listed.

</think>

<answer>

{"evidence_pages": [3, 4, 5], "answer": "11"} **Wrong Answer**

</answer>

Question

What date is mentioned on page three?

Ground Truth

Answer: 1916-02-17 Evidence_pages: [3] Evidence_sources: ['Chart']

<think>

Question Analysis

The user is asking for a specific piece of information: a date. The question specifies that this date should be found on "page three". My task is to locate any date mentioned within the provided document content, which corresponds to page 3.

Evidence Localization

I am reviewing the provided OCR text, which is confirmed to be from page 3. Near the top of the page, below the title and author, there is a line that reads: "Read 17th February 1916." This is a clear and explicit date.

Reasoning Process

The question asks for the date mentioned on page three. The evidence localization step identified "17th February 1916" as the only date present on the provided page content. Therefore, this is the direct answer to the question.

</think>

<answer>

{"evidence_pages": [3], "answer": "17th February 1916"} **Answer correct but formatting mismatch**

</answer>

Figure A5. Failure cases in ALR CoT Data Distillation.

Case Study

Question

Assume that in a survey conducted after July 2020, the percentage of Americans who believe China has done a "good" job dealing with the coronavirus outbreak increased by 10 percentage points, then what percentage of Americans would believe China has done a "bad" job (assuming the percentage of all the other options stays the same)?



Ground Truth

Answer: 54%

Evidence_pages: [5]

Evidence_sources: ['Chart']

China has done a __ job dealing with the coronavirus outbreak



Baseline

Localization failure

Based on ... if the percentage of ... a "good" job ... increased by 10 percentage points, we can calculate the new percentage for those who believe China has done a "bad" job. Let's assume the original percentage of ... a "good" job was (x). After the increase, it becomes (x + 10) ... Therefore, the percentage of Americans who believe China has done a "bad" job is: [100% - (x + 10) = 90% - x]. Without knowing the exact original value of (x), **we cannot provide a specific numerical answer**. However, if you provide ... I can calculate ...

Fine-tuned on short answer data

Wrong Answer & Missing Reasoning Process

31.

Vanilla CoT

Correct localization & Incorrect reasoning

<think>

1. Analyze ...; 2. Scan the document: ... 3. Locate the relevant chart: I see a **bar chart titled "Most Americans see China as having dealt poorly with COVID-19"** shows ...; 4. Examine the chart's data: The chart has two categories: "Bad" and "Good." ...; 5. Extract the specific numbers: The question states ... increased by 10 percentage points. Looking at the chart, the "Good" category for the "Total" group is **31%**. ... the original percentage as (x). The problem says (x + 10 = 31). **Solving for (x): (x = 31 - 10 = 21)**. So, the original percentage ... "good" job was **21%**. 6. Calculate the percentage who believe China has done a "bad" job: The total must be 100%. If 21% believe it has been good, then (100% - 21% = **79%**) believe it has been bad...

</think>

<answer>**79%**</answer>

DocSeeker-SFT

Correct localization Incomplete reasoning

<think>

{Question Analysis}: ... {Evidence Localization}: I need to find the initial percentages for "Bad" and "Good" jobs from the provided document.- The document states: ... This refers to blame, not ... However, there is a chart titled "... " which shows: ... **-Bad: 64% - Good: 31%**. {Reasoning Process}: ... From the evidence, the initial percentage for "Good" was 31%. An increase of 10 percentage points means the new percentage for "Good" is 31% + 10% = 41%. Since the total must sum to 100%, the percentage for "Bad" will be 100% - 41% = **59%** ...

</think>

<answer>{"evidence_pages": [5], "answer": "output: **59**"}</answer>

DocSeeker

Correct localization Correct reasoning

<think>

{Question Analysis}: ... {Evidence Localization}:... **the "Bad" shows 64% and the "Good" shows 31% ...** {Reasoning Process}: The question states ... increased by 10 percentage points. From the evidence, the initial percentage for "Good" is 31% ... the new percentage for "Good" is 31% + 10% = 41%. Since the total percentage of all responses must sum to 100%, the percentage for "Bad" will be 100% - 41% = 59%. Therefore, if the "Good" percentage increases by 10 points, the "Bad" percentage will decrease by 10 points from its original value of 64%. The original "Bad" percentage was 64%, so the new "Bad" percentage is **54%**.

</think>

<answer>{"evidence_pages": [5], "answer": "output: **54**"}</answer>

Figure A6. Qualitative Comparison of Reasoning Paradigms.

Table A4. Detailed Configurations of Compared Methods on Long Document Understanding.

| Method | Model | OCR | Retrieval Config | | Backbone | Trained on Dataset | | |
|-------------------|--------------------|-----|-----------------------|-------|-------------------|--------------------|----------|----------|
| | | | Retriever | Top-k | | DUDE | MPDocVQA | SlideVQA |
| RAG | CREAM [42] | ✓ | bge-large [49] | 3 | Pix2Struct/LLaMa2 | ✓ | ✓ | × |
| | M3DocRAG [43] | × | Colpali [18] | 4 | Qwen2-VL | × | × | × |
| | Vis-RAG [17] | × | VisRAG-Ret [17] | 3 | MiniCPM-V 2.6 | × | × | × |
| | SV-RAG [32] | × | SV-RAG-InternVL2 [32] | 5 | SV-RAG-InternVL2 | × | × | ✓ |
| | VDocRAG [33] | × | VDocRetriever [33] | 3 | VDocGenerator | ✓ | × | × |
| End-to-End | HiVT5 [1] | ✓ | - | - | DiT/T5 | × | ✓ | × |
| | mPLUG-DocOwl2 [10] | × | - | - | ViT/LLaMa+MAM | ✓ | ✓ | × |
| | Docopilot [44] | × | - | - | InternVL2 | ✓ | ✓ | × |
| | DocVLM [45] | ✓ | - | - | Qwen2-VL | × | × | × |
| | InternVL3 [9] | × | - | - | InternViT/Qwen2.5 | × | × | × |

Table A5. Comparison of method architectures and attributes, where "OCR" and "Retriever" denote dependency on external modules, "Training-free" indicates the absence of additional fine-tuning on multi-page documents, "Page-level Reasoning" refers to processing at page granularity, and "Evidence Localization" marks the capability for explicit source grounding.

| Method | OCR | Retriever | Backbone | Training-free | Page-level Reasoning | Evidence Localization |
|--------------------|-----|-----------------------|-----------------------|---------------|----------------------|-----------------------|
| <i>OCR-based</i> | | | | | | |
| Longformer [29] | ✓ | × | RoBERTa | ✓ | × | × |
| DocLLM [50] | ✓ | × | Falcon/LLaMa2 | ✓ | × | × |
| LayTokenLLM [30] | ✓ | × | LLaMa3/LLaMa2/Qwen1.5 | × | × | × |
| <i>End-to-End</i> | | | | | | |
| HiVT5 [1] | ✓ | × | DiT/T5 | × | × | × |
| mPLUG-DocOwl2 [10] | × | × | ViT/LLaMa+MAM | × | ✓ | × |
| Docopilot [44] | × | × | InternVL2 | × | ✓ | × |
| DocVLM [45] | ✓ | × | Qwen2-VL | ✓ | × | × |
| InternVL3 [9] | × | × | InternViT/Qwen2.5 | ✓ | ✓ | × |
| Qwen2.5VL [8] | × | × | QwenViT/Qwen2.5 | ✓ | ✓ | × |
| <i>RAG</i> | | | | | | |
| PDF-WuKong [20] | ✓ | bge-m3 [49] | XComposer2-4KHD | × | × | ✓ |
| CREAM [42] | ✓ | bge-large [49] | Pix2Struct/LLaMa2 | × | × | ✓ |
| M3DocRAG [43] | × | Colpali [18] | Qwen2-VL | ✓ | ✓ | ✓ |
| Vis-RAG [17] | × | VisRAG-Ret [17] | MiniCPM-V 2.6 | ✓ | ✓ | ✓ |
| SV-RAG [32] | × | SV-RAG-InternVL2 [32] | SV-RAG-InternVL2 | × | ✓ | ✓ |
| VDocRAG [33] | × | VDocRetriever [33] | VDocGenerator | × | ✓ | ✓ |

F.2. Efficacy of the Two-Stage Training Framework

Figure A7 illustrates the evolution of model capabilities within the two-stage training framework. After SFT on ALR CoT data, DocSeeker-SFT successfully acquires the structural form of the ALR paradigm, strictly adhering to the "Analysis-Localization-Reasoning" workflow; however, it exhibits unstable grounding capabilities, evidenced by a incorrect page that results in an erroneous conclusion. This indicates that while SFT facilitates the imitation of reasoning patterns, it is insufficient for ensuring factual precision. The subsequent introduction of Evidence-aware GRPO effectively rectifies this limitation. Guided by

localization-specific reward signals, the model achieves precise evidence grounding while retaining the structured reasoning path, demonstrating that the reinforcement learning stage is critical for optimizing localization accuracy and enhancing overall performance in long-document understanding.


G. Related Work

To provide a more intuitive comparison of distinct technical approaches in the field of long-document understanding, we summarize related studies in Table A5. These methods can be primarily categorized into the following three paradigms:


Case Study

Question


What position is the man with the red shirt in the figure on the last page, directly answer 'bottom', 'middle' or 'top'?




Page 1



Page 40



Page 76



Page 77

Ground Truth

Answer: bottom
Evidence_pages: [77]
Evidence_sources: ['Figure']

Baseline

Response:
 The image shows a group of people posing for a photo. The text “Thank u very much.” appears at the top of the image, which is likely a message of gratitude from one person to another. This could be

If you have any specific questions about the image or need further details, **feel free to ask!**

❌ **Fail to answer**

Answer: Fail to answer
Score: 0.0

Fine-tuned on short answer data

Response:
 {Question Analysis}
 The user is asking ... This requires...
 {Evidence Localization}
 ... is **page 71** ...see a group of people. One individual in the front row, ...
 ❌ **Incorrect localization**
 {Reasoning Process}
 This individual is positioned centrally in the front row ... suggests a prominent or middle position. Therefore... is **in the middle**.
 ❌ **Incorrect reasoning**
 {Answer}
 "evidence_pages": [71]
 "answer": "middle"
Score: 0.0

DocSeeker

Response:
 {Question Analysis}
 The user is asking ... This requires...
 {Evidence Localization}
 ... **page 77**...see a photo of several men. In the front row, ...a man wearing ... is positioned centrally in the front row.
 ✅ **Precise localization**
 {Reasoning Process}
 The man wearing the red shirt is located in the lower half ... specifically in the front row... the **lowest visible row** of people ...Therefore, ... is **"bottom"**
 ✅ **Correct reasoning**
 {Answer}
 "evidence_pages": [77]
 "answer": "bottom"
Score: 1.0

Figure A7. Performance Comparison across Different Training Stages.

1) OCR-based Methods: These approaches, such as Lay-TokenLLM [30] and DocLLM [50], rely on external OCR engines to extract textual information from documents, which then serves as input for the model.

2) RAG-based Methods. To address the limitations of MLLMs in processing ultra-long documents, these methods, such as Vis-RAG [17] and VDocRAG [33], introduce

a retriever to select the Top-*k* most relevant pages for subsequent reasoning.

3) End-to-End Methods. These approaches, such as mPLUG-DocOwl2 [10] and InternVL3 [9], directly encode document images into visual tokens for processing, thereby preserving complete visual features.