

OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation

Supplementary Material

I. Instruction Prompts

Q&A Generation Prompt Template. The template is shown in Tab. S9. Following [53, 58], we instruct GPT-4o to generate questions with clear entities and require three levels of difficulty for question diversity.

RAG Generation Stage Prompt Template. The prompt template for LLMs and VLMs with text-only input is shown in Tab. S12.

Vision-Language Models OCR Prompt Template. We tune the prompt for the best performance of VLMs OCR, by comparing simple and detailed instructions as shown in Tab. S13. Results in Tab. S14 indicate that the detailed prompt consistently performs better across all evaluations, so it is used by default.

II. Benchmark Construction Details

II.1. Document details

We curate a dataset of 1,261 PDFs spanning 8,561 pages, with 3,596 pages designated for Q&A generation and the remainder forming part of the knowledge base. These PDFs are sourced from DUDE [46], OmniDocBench [35], FinanceBench [22], CUAD [18], GNHK [24], and public resources, including Arxiv⁴, ManualsLib⁵, LibreTexts⁶.

DUDE: We extract documents from the validation and test splits of DUDE, applying manual screening based on our criteria Fig. 1 to exclude samples infeasible for structured data parsing and classify each of them into 7 domains. We finally selected 450 PDFs with 4,058 images from 2,069 PDF candidates.

OmniDocBench: OmniDocBench [35] features span-level annotations and presents challenges for OCR due to its multilingual, high-resolution with dense text and handwritten content. We select all newspaper documents and manually review textbook-related samples, eliminating those with low knowledge density that hinder meaningful Q&A generation. This process yields 289 PDFs.

FinanceBench: Following prior observations [33], both DUDE [46] and FinanceBench [22] contain diverse document types. From FinanceBench, we randomly sample 10 PDFs characterized by large, complex tables and charts.

CUAD: We randomly select 65 PDFs to supplement the documents in law domains, which all have high text density.

GNHK: GNHK consists of handwritten documents. We manually assess and remove those with low knowledge density, finalizing a selection of 172 PDFs.

Each document is manually reviewed by primary authors to ensure its availability for academic use. Detailed domain statistics are shown in Tab. S1

Domains	PDFs	Pages	Pages with Q&As
Law	95	1187	1143
Finance	65	2133	1359
Textbook	504	678	1126
Manual	87	1724	1155
Newspaper	279	487	1202
Academic	85	1011	1181
Administration	146	1341	1332
Total	1261	8561	8498

Table S1. Document statistics of each domain

II.2. Ground truth structured data annotation

We annotate the ground truth structured data using Mathpix Markdown format, where tables and formulas are represented in LaTeX. Chart data is extracted in LaTeX table format, with charts lacking clear numeric values in figure filtered out. For images in documents, any parsable text is retained as plain text in the corresponding section. To ensure high-quality annotations, we first use Mathpix to pre-annotate all PDFs. Finally, the primary authors employ Mathpix Markdown previews⁷ to render structured data into PDFs, manually review and correct pre-annotated results.

II.3. Document with challenging attributes

Although existing RAG document benchmarks have gathered PDFs from different domains [10, 16, 20], they often ignore the challenges posed by OCR. To address this, we construct a benchmark that explicitly incorporates documents with challenging attributes. We define nine key attributes: structured data (tables, formulas, charts), complex layouts, handwritten content, distortions, scanned PDFs, dense text, and multilingual content. Structured data, dense text (exceeding 770 tokens), and multilingual pages are classified based on the annotated ground truth structured data. A document is considered to have a complex layout if its layout

⁴<https://arxiv.org>

⁵<https://www.manualslib.com/>

⁶<https://libretexts.org/>

⁷<https://github.com/Mathpix/vscode-mathpix-markdown>

detection yields more than 20 bounding boxes. Distorted, scanned, and handwritten documents are identified during manual checks.

II.4. Q&A generation

To generate high-quality Q&A pairs covering diverse tasks and evidence sources, we define multiple prompts for each task, as detailed Tabs. S9 to S11. For Chinese questions, we provide the same set of templates in Chinese to ensure that the model generates Chinese responses. **Q&A with different evidence sources.** For Q&A generation with evidence sourced from plaintiff text, table, formula and chart, we extract relevant pages from the ground truth structured data and use GPT-4o to generate Q&A pairs grounded in the corresponding evidence via tailored prompts. For Q&A related to reading order, we leverage MinerU [48], the leading model for reading order recognition [35], to identify the reading order and bounding box of paragraphs in each document. When working with documents from OmniDocBench [35], we directly use the ground truth reading order from its annotations. We verify the layout detection and reading order predictions, selecting paragraph pairs that meet one of the following criteria:

- Adjacent paragraphs in reading order whose bounding boxes are not vertically aligned.
- Paragraphs separated by multimodal document elements (e.g., block formulas, tables, or images).

We then randomly sample 1,500 candidate matches, manually correcting approximately 20% where MinerU’s predictions are inaccurate. We then prompt GPT-4o to generate Q&A pairs using the prompts in Tab. S10. We find that this simple prompting-based strategy can effectively generate questions with diverse evidence sources, with over 90% correctly aligned with their evidence source in our Q&A verification process.

Q&A with different tasks. To generate both understanding and reasoning questions, we apply the corresponding prompts from Tab. S10. For multi-page Q&A generation, we employ two different approaches to generate Q&A candidates: (1) Combine questions from two single-page Q&As that mention the same entity. (2) Generating multi-page questions from two paragraphs on different pages that reference the same entity. Specifically, we use spaCy [19] for named entity recognition in both single-page Q&As and document paragraphs. We then filter out candidate pairs, including: (1) Single-page Q&A pairs where the entity in one answer appears in another question. (2) Paragraph pairs that share the same entity. We finally utilize the prompts in Tab. S11 to generate multi-page questions. However, despite the many optimizations of the prompt and generation strategies we tried, GPT-4o sometimes produces Q&A pairs that are either answerable with a single paragraph or simply concatenate two single-page questions while maintaining separate evi-

dence sources instead of high-quality and realistic multi-page Q&As. To address these limitations, we develop a comprehensive filtering process to ensure the quality of multi-page Q&As, as detailed in Sec. II.5.

II.5. Q&A verification.

We verify Q&A quality based on three criteria: (1) Compatibility with realistic RAG applications, (2) faithfulness to task definition, and (3) correctness. Below, we detail our approach for each aspect.

Compatibility with Realistic RAG Applications. To assess context dependence, we identify key patterns from existing context-dependent questions and apply the following heuristics:

- Questions lacking an explicit entity name.
- Questions containing more than one ambiguous pronouns (e.g., "he," "she," "it," "they", "this", "that").
- Questions featuring phrases such as "in the document" or "according to the document."

These rules filter most context-dependent questions. We then refine the selection using prompts in VisRAG [56] and DeepSeek-V3 to further distinguish context-dependent questions from the remaining set. Additionally, we use GPT-4o to exclude questions answerable without retrieval by instructing it to respond without providing evidence context across both single-page and multi-page Q&As.

Faithfulness to Task Definition. Based on the Q&A verification prompts in [12], we use the prompts in Tab. S15 to assess faithfulness using DeepSeek-V3. To verify the validity of evidence sources, we locate them in the original ground truth structured data and ensure they originate from the correct corresponding LaTeX code environments. For the multi-page and reading-order questions, we employ GPT-4o to generate three responses: (1) without context, (2) with context A, and (3) with context B. If any response yields a correct answer, the question is excluded, ensuring that only truly multi-page or reading-order-related questions remain.

Correctness. To guarantee each Q&A has a unique and correct answer supported by its evidence context, we provide oracle evidence and sample GPT-4o’s response 10 times. We apply a best-of-N strategy to determine the final answer, which must match the ground truth. Q&As with fewer than three consistent correct responses are also excluded.

Our filtering pipeline underwent two iterations of refinement. In each round, we randomly sample 100 Q&As to verify the filtering results adherence to our criteria. Finally, to mitigate false positives, we manually reviewed all remaining questions, yielding 8,498 high-quality Q&As from an initial pool of 15,317 candidates.

III. OCR Noise Introduction

III.1. Rules for Formatting Noise introduction

To introduce *Formatting Noise*, we define a perturbation rate r to control its extent. In order to match the level of *Semantic Noise* (measured by similar edit distance), we set the $r = \{0.1, 0.3, 0.6\}$, indicating the three levels of perturbation: mild, moderate, and severe. Based on the *Formatting Noise* in the existing OCR results, we formulate the following rules to perturb plain text, tables, and formulas, respectively.

III.1.1. Plain text

Text Style: Given the plain text content of the ground truth, we randomly divide it into a sequence where each item consists of 2 to 5 words, select target items based on r , and apply one of the following operations as perturbations.

- **Bold:** Enclose the selected text in `**` or `\textbf{}`.
- **Italic:** Enclose the selected text in `*` or `\textit{}`.
- **Underline:** Enclose the selected text in `_` or `\underline{}`.

Title Formatting: We identify short sentences that end with a full stop and have no more than 5 words as potential headings. We randomly pick them according to r and add one of level 1 to level 3 title controls in Markdown (`#`) or LaTeX (`\section{}`) to make new titles.

Paragraph: To simulate the line breaks that exist in PDFs, we randomly insert `\n` at word intervals based on r .

III.1.2. Formula

Formula Conversion: Randomly convert the inline formula into block formula and vice versa at rate r .

Extraneous Elements: We first randomly select the target formulas based on r . Subsequently, for each target formula, we randomly insert 1 to 5 meaningless markers in its symbol gaps, including `\`, `\quad`, `\qquad`, `\;`, `\:`.

Equivalent Symbols: For each formula, we replace the following equivalent symbols with probability r :

- **bold:** `\mathbf{}`, `\boldsymbol{}`.
- **cursive:** `\mathbb{}`, `\pmb{}`, `\mathrsfs{}`, `\euscript{}`, `\mathcal{}`.
- **unicode:** (`\sigma`, `\u03A3`), etc⁸.

III.1.3. Table

Row and Column Lines: For each line and column, randomly insert `\hline` or `\cline` with probability r .

Cell Content: For each cell content, randomly apply above rules for plain text or formula with probability r .

III.2. Rules for Semantic Noise introduction

In order to construct perturbed document images that conform to the realistic distribution of naturally distorted docu-

OCR	Avg. Counts
MinerU	35.0
GOT	45.7
Nougat	63.2
F-minor	37.9
F-moderate	42.2
F-severe	56.3

Table S2. Counts of *Formatting Noise*. The counts of *Formatting Noise* we add (F-minor, F-moderate, F-severe) is approximately the distribution of the counts of *Formatting Noise* for MinerU, GOT and Nougat.

ments, we use a cross-validated process involving multiple annotators. We finally identify 8 strategies from [5] as follows:

- **Background Addition:** We collect 15 background images of real paper textures and blend them with original images at an 80:20 ratio.
- **Salt-and-Pepper Noise:** Randomly replace 1% of the image pixels with white ("salt" noise) and black ("pepper") pixels.
- **Dirty Rollers:** Add random rollers with thickness between 1 and 3 pixels, a line addition probability of 0.05 per pixel row or column.
- **Random Rotation:** Apply a random rotation of -3° and $+3^\circ$.
- **Binarization:** Utilize the Augraphy⁹ to simulate effects such as dilation, erosion, and letterpress printing.
- **Warping:** Apply geometric transformations and folding effects via Augraphy to mimic paper creases.
- **Shadows:** Apply light gradients and shadow cast from Augraphy to simulate shadows that occur when a document is taken.
- **Blur via Point Spread Function:** Generated 100 PSF kernels and randomly applied one to the document.

We classify above distortions into two categories: (1) weak distortions: These preserve text clarity and include background addition, binarization, minor rotation, and PSF-based blurring. (2) strong distortions: These degrade readability, causing blurriness and font warping. They include salt-and-pepper noise, dirty rollers, warping, and shadow effects. To simulate varying levels of document distortion, we apply the above strategies in three ways:

- Apply a weak distortion per page.
- Apply a strong distortion per page.
- Apply two randomly selected distortions per page.

We generate three document image datasets with varying noise levels and parse structured data using MinerU, GOT, and Qwen2.5-VL, resulting in 9 perturbed datasets. The examples of distorted documents are shown in Fig. S1. The distribution of introduced *Semantic Noise* is illustrated

⁸Full lists are drawn from <https://raw.githubusercontent.com/w3c/xml-entities/refs/heads/gh-pages/unicode.xml>

⁹<https://github.com/sparkfish/augraphy>

OCR		SN		
	TXT	FOR	TAB	FN
MinerU	40.3%	78%/34%	79%/58%	10.9
Qwen2.5-VL	31.6%	46%/25%	75%/60%	16.4

Table S3. SN: ratio of matching blocks with textual/structural errors. FN: average redundant formatting commands per page.

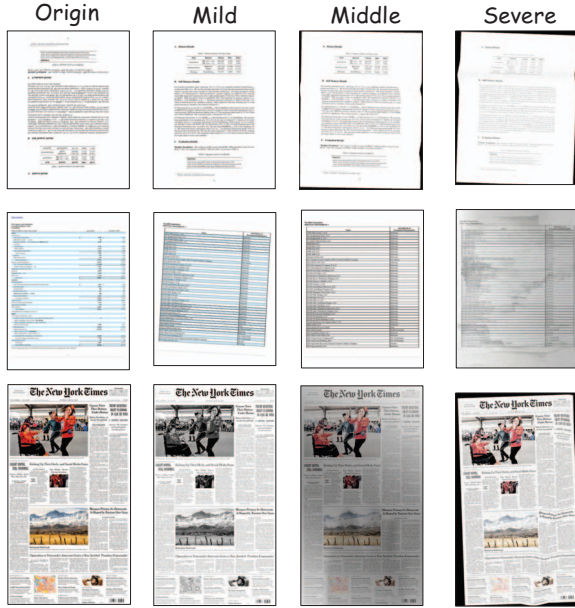


Figure S1. Cases of distorted documents.

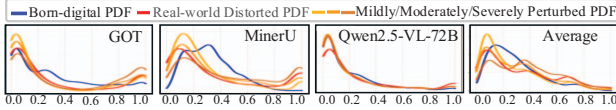


Figure S2. Distribution of *Semantic Noise*. X-axis denotes edit distance. Mild/moderate/severe perturbation is based on born-digital PDFs.

in Fig. S2. In most cases, the distributions of our perturbed PDFs align with those of real-world distorted PDFs, validating the realism of our method. In Sec. 4.3, we evaluate RAG performance on these datasets, reporting the average results for each noise level.

Ratio of OCR noise in real-world OCR results. To illustrate the frequency of OCR noise in real-world OCR results, We match corresponding TXT/FOR/TAB blocks, which includes ~130 tokens each, and show the ratio of *Semantic Noise* and *Formatting Noise* in Tab. S3.

	TXT↑	TAB↑	FOR↑	CHA↑	RO↑	ALL↑
GT	57.7	41.7	41.8	39.8	29.8	46.8
MinerU	<u>51.2</u>	33.7	<u>32.1</u>	10.7	29.8	37.9
Qwen2.5-VL	54.1	<u>38.9</u>	34.5	<u>22.7</u>	13.3	40.6
VisRAG	50.7	40.3	32.0	30.6	<u>15.6</u>	<u>40.2</u>

Table S4. Performance of VisRAG and OCR-based RAG. We use Qwen2-VL-7B as the generator for fair comparison.

IV. Additional Experimental Results

IV.1. Experimental details

For MinerU, we use version 0.9.2¹⁰ by default. For Marker, version 0.2.17¹¹ is employed. For Nougat, we utilize its 0.1.0-base model (350M). All prompt templates can be found in Sec. I.

For all LLMs and VLMs, we set the temperature to 0 with `do_sample=False` by default for reproducibility.

IV.2. Sim2Real GAP

As the questions posed by human users could have far more diversity in styles than LLM generated Q&As. We randomly pick 100 Q&As and manually rewrite questions for comparison. The performance before and after rewriting is: 27.2/23.2(GT), 20.7/18.0(MinerU), 12.8/12.9(GOT), and 23.1/20.0(Qwen2.5-VL). Although performance degrade, the conclusions about different OCR solutions still hold, as question styles may primarily be associated with models' ability to understand instructions.

IV.3. Multimodal RAG

We compare VisRAG with OCR-based RAG, using Qwen2-VL-7B as the generator for fair comparison. The results are shown in Tab. S4. VisRAG achieves competitive results on multimodal element-related Q&As (e.g. table and chart), but underperforms on TXT and RO (e.g. high-resolution newspapers), exhibiting similar failure modes to Qwen2.5-VL.

IV.4. Effectiveness of robust generator

We employ Ext2Gen-8B-R2 [41] and show its performance in Tab. S5. Ext2Gen-8B-R2 consistently improves performance. Although it is based on Llama3.1-8B, its performance on Azure remains stable, reinforcing that stronger models exhibit greater robustness to formatting noise. This further supports our conclusion that stronger models are more robust to formatting noise. However, the performance gap between the best OCR (Azure) and GT also increases

¹⁰https://github.com/opendatalab/MinerU/releases/tag/magic_pdf-0.9.2-released

¹¹<https://github.com/VikParuchuri/marker/releases/tag/v0.2.17>

OCR	E.D.	TXT \uparrow	TAB \uparrow	FOR \uparrow	CHA \uparrow	RO \uparrow	ALL \uparrow	$\Delta(\text{ALL})$
Generator: Qwen2-7B/Llama3.1-8B								
GT	-	46.7/43.1	31.8/37.4	27.6/28.4	31.1/34.7	23.7/13.7	36.2/35.9	-
MinerU	0.24	42.2/37.8	27.0/30.0	23.5/22.5	8.9/9.7	23.0/12.5	30.5/28.5	-5.7/-7.4
Qwen2.5-VL	0.18	42.5/38.6	29.1/33.1	26.1/26.1	18.5/19.6	10.9/6.7	31.5/30.7	-4.7/-5.2
Azure	0.17	45.5/29.6	30.7/25.4	23.3/21.9	19.1/11.0	23.5/11.5	33.8/24.0	-2.4/-11.9
Generator: Ext2Gen-8B-R2								
GT	-	56.3	45.4	40.7	38.9	27.4	46.8	-
MinerU	0.24	49.7	36.4	30.5	10.8	25.8	37.6	-9.2
Qwen2.5-VL	0.18	52.7	41.2	34.8	24.6	12.9	40.9	-5.9
Azure	0.17	55.1	41.9	32.4	23.4	26.0	42.8	-4.0

Table S5. Experiments of Azure and Ext2Gen.

by 1.6 compared to Qwen2-7B, indicating that OCR quality becomes a bottleneck and leaves a room for improvement.

IV.5. Commercial OCR

We evaluate Azure OCR in Tab. S5 and observe the following: With powerful generators (Qwen2-7B and Ext2Gen-8b-R2), Azure yields the best performance, though there remains a gap of up to 4.0 compared to GT. But, when using Llama3.1-8B, performance drops significantly, even worse than MinerU. Our manual check suggests this may be due to custom formatting tags in Azure’s outputs, affecting Llama3.1-8B’s generation.

IV.6. Details in different domains

Tab. S6, Tab. S7 and Tab. S8 shows the performance of different OCR solution on different domains respectively.

V. Case Study

Fig. S3 to Fig. S12 show some cases of GOT, MinerU, and Qwen2.5VL-72B on OHRbench. For each case, we indicate the evidence source and answer, giving the OCR result of different models and the responses at the retrieval and generation stages.

Domain	GT	MinerU	Marker	GOT	Nougat	Qwen2.5-VL	InternVL2.5
Law	81.2	71.0	77.1	62.1	69.0	76.4	69.6
Finance	59.7	36.4	45.0	30.4	25.8	47.9	47.1
Textbook	73.2	43.8	49.6	48.8	37.1	58.3	55.0
Manual	79.1	60.4	68.6	58.9	47.8	71.3	70.2
Newspaper	40.5	31.3	34.0	12.4	10.6	27.7	18.4
Academic	75.1	50.3	55.2	50.2	45.0	61.1	57.1
Administration	82.2	59.4	68.3	57.7	52.7	73.1	73.8
All	70.0	50.1	56.6	45.4	40.8	59.2	55.8

Table S6. Retrieval performance across different domains.

Domain	GT	MinerU	Marker	GOT	Nougat	Qwen2.5-VL	InternVL2.5
Law	56.9	53.4	54.4	43.3	48.8	53.9	50.9
Finance	43.1	30.1	29.5	19.7	17.7	35.9	36.8
Textbook	37.6	25.9	28.2	24.8	16.8	29.1	29.1
Manual	50.2	45.3	46.1	41.3	34.3	48.7	47.7
Newspaper	35.0	33.7	31.6	9.5	8.4	19.6	11.7
Academic	38.3	29.5	27.9	25.3	24.8	33.2	31.3
Administration	46.4	35.7	37.7	32.2	29.2	42.7	42.9
All	43.9	36.1	36.3	27.8	25.5	37.5	35.8

Table S7. Generation performance across different domains.

Domain	GT	MinerU	Marker	GOT	Nougat	Qwen2.5-VL	InternVL2.5
Law	49.6	48.1	48.1	41.1	43.9	47.2	44.9
Finance	27.2	19.4	20.1	15.1	13.1	22.9	22.8
Textbook	30.5	20.9	22.5	21.0	15.7	23.8	23.5
Manual	44.4	38.1	39.8	36.0	30.7	42.3	41.6
News	29.0	25.6	24.7	8.3	5.6	17.4	11.0
Academic	31.9	25.6	24.1	22.8	21.2	27.6	26.4
Administration	41.0	30.9	32.7	29.2	26.6	37.3	37.5
All	36.1	29.5	30.0	24.6	22.2	31.1	29.6

Table S8. Overall performance across different domains.

System:

You are an AI specialized in generating QAs from documents. Your mission is to analyze the document, follow the instructions, and generate RAG-style question-answer pairs based on the document.

RAG-style refers to a question that needs to be answered by retrieving relevant context from an external document based on the question, so the question **MUST** obey the following criteria:

1. Question should represent a plausible inquiry that a person (who has not seen the page) might ask about the information uniquely presented on this page. The questions should not reference this specific page directly (by page number, pointing to a specific paragraph or figure, and never refer to the document using phrases like 'in the document'), nor should they quote the text verbatim. They should use natural language reflecting how someone might inquire about the page's content without direct access.
2. Question must contain all information and context/background necessary to answer without the document. Do not include phrases like "according to the document" in the question.
3. Question must not contain any ambiguous references, such as 'he', 'she', 'it', 'the report', 'the paper', and 'the document'. You **MUST** use their complete names.

User:

Your task is to generate several RAG-style question-answer pairs with different levels of difficulty and evidence sources. {detailed_task_description}.

You **MUST** obey the following criteria:

- The question **MUST** be detailed and be based explicitly on information in the document.
- The question **MUST** include at least one entity.
- The context sentence the question is based on **MUST** include the name of the entity. For example, an unacceptable context is "He won a bronze medal in the 4 × 100 m relay". An acceptable context is "Nils Sandström was a Swedish sprinter who competed at the 1920 Summer Olympics."
- The answer form should be as diverse as possible, including [Yes/No, Numeric, String, List].
- {additional_task_criteria}

If there are no possible questions that meet these criteria, return 'None' as the question. Output the question in JSON format.

{qa_examples}

<document>{document}</document>

Table S9. Q&A Generation Prompt

Structure data task:

In the given documents, the chart elements are all enclosed within `<chart>` `</chart>` tags and illustrated in LaTeX table format. Pay attention to the difference between them and tabular data, as tabular data is not enclosed by `<chart>` `</chart>` tags. **# This paragraph is only used for chart data.**

In order to generate this type of question-answer pairs, first, you need to read the given document, identify the table/formula/chart elements within it, and use them as the evidence context. The evidence context can be a single paragraph for single-hop questions, or several related paragraphs for generating multi-hop questions that require reasoning. After that, you need to generate questions and corresponding answers based on them.

Reading order task:

Your task is to generate RAG-style question-answer pairs from the given two documents. In order to generate this type of question-answer pairs, first, you need to read the given two documents (A, B), identify the text sharing the same entities, and design a question-answer pair based on the contents of both documents A and B. If it is based on the message of document A or document B alone, it cannot be answered.

Understanding task:

You should generate question-answering pairs that require the responder to extract information from documents. The answer should be able to find directly in the documents without any reasoning.

Reasoning task:

You should generate question-answering pairs that require responder to reason before answering, such as calculations, comparisons, finding the maximum and minimum, or integration information from different parts of the documents. The answer should not be able to be found directly in the documents.

Table S10. Detailed description used to generate Q&A pairs for different tasks.

Multi-page Q&A from single-page question:

Your mission is to generate RAG-style combined questions from two questions that have the same entity.

When generating a combined question, there are some criteria you should follow:

- The answer to the combined question should be the same as the answer2.
- It must combine the answer1 to question1 to answer the combined questions. This means that, to answer the combined question, a responder must first deduce the part of the combined question that refers to the answer1, and then proceed to answer the combined question based on that answer.
- You cannot include the answer to question 1 in the combined question.

{combined_qa_examples}

Based on the above 3 examples, provide a combined question for the following case. If you find it is hard to create such a combined question, output None as the answer. Enclose the combined question within <answer></answer>:

question1: {q1}

answer1: {a1}

question2: {q2}

answer2: {a2}

Multi-page Q&A from different paragraphs:

Your task is to generate RAG-style question-answer pairs from the given two documents and entity names. The entity names appear in both documents, and you need to use them as a bridge to generate the RAG-style question-answer pairs that need to be answered by combining information from both documents.

To generate the question-answer pairs, first, you need to read the given two documents (A, B) and the entity names, find paragraphs related to them, use the paragraphs as evidence context, and design a question-answer pair based on the evidence context from the two documents.

Table S11. Detailed description used to generate multi-page Q&A pairs from both single-page questions and different paragraphs sharing same entities.

System:

You are an expert, you have been provided with a question and documents retrieved based on that question. Your task is to search the content and answer these questions using the retrieved information.

You **MUST** answer the questions briefly with one or two words or very short sentences, devoid of additional elaborations.

Write the answers within <response></response>.

User:

Question: {question}

Retrieved Documents: {retrieved_documents}

Table S12. LLMs prompt for RAG generation

Simple Prompt:

Please do OCR on the image and give all the text content in markdown format. The formulas should be wrapped in \$\$\$. The table and charts should be parsed in LaTeX format. Only output the OCR results without any extra explanations or comments.

Table S13. Simple prompt for VLMs OCR

Detailed Prompt:

You are a powerful OCR assistant tasked with converting PDF images to the Markdown format. You MUST obey the following criteria:

1. Plain text processing:

- Accurately recognize all text content in the PDF image without guessing or inferring.
- Precisely recognize all text in the PDF image without making assumptions in the Markdown format.
- Maintain the original document structure, including headings, paragraphs, lists, etc.

2. Formula Processing:

- Convert all formulas to LaTeX.
- Enclose inline formulas with $$. For example: This is an inline formula $E = mc^2$.$
- Enclose block formulas with $$. For example:
$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$
.$

3. Table Processing:

- Convert all tables to LaTeX format.
- Enclose the tabular data with \begin{table} \end{table} .

3. Chart Processing:

- Convert all Charts to LaTeX format.
- Enclose the chart data in tabular with \begin{table} \end{table} .

4. Figure Handling:

- Ignore figures from the PDF image; do not describe or convert images.

5. Output Format:

- Ensure the Markdown output has a clear structure with appropriate line breaks.
- Maintain the original layout and format as closely as possible.

Please strictly follow these guidelines to ensure accuracy and consistency in the conversion. Your task is to accurately convert the content of the PDF image using these format requirements without adding any extra explanations or comments.

Table S14. Complex prompt for VLMs OCR

System:

You are an AI specialized in document question-answering verification. Your mission is to analyze the given question-answering pairs and follow the instructions. Your response must be true and accurate, and no additional content should be output.

1. Question type check

Dose the question match the task description: {detailed_task_description}

Make sure the question meets the required task context.

2. Evidence relevance Check

Dose the provided evidence context relate to the question provided? Does the answer accurately reflect the information in the evidence context? Ensure the question is formulated based on information explicitly stated. The question should not introduce concepts unrelated to the document's content.

3. Clarity and Precision

Is the question clear and unambiguous? And is the answer concise and precise? Ensure the language is straightforward and easily understandable, and avoid complex phrasing that may confuse the reader. The intention of the question and answer pair must be clear and direct, avoiding verbosity and unnecessary detail. Ensure the answer fully addresses the question without omitting crucial information.

{qas}

Table S15. Q&A Verification Prompt

Evidence Source: Text

who had been vaccinated three months before death, from the arm of a healthy child. Three other children vaccinated at the same time, from the same source, took no hurt. On the eighth day after vaccination, a papular and vesicular rash appeared over the trunk, which rapidly assumed a sloughing character. The eruption was at first taken for small pox, and when death took place, a fortnight later, an inquest was held on the case, for it was then thought to be syphilis. But Mr. HUTCHINSON pointed out that its evolution as well as its character were not those of syphilitic infection, and he considered it to be a true case of vaccine passing on to a gangrenous condition—a condition he had sometimes observed to take place in variola. The vaccine marks on the arm were natural.—MR. JONATHAN HUTCHINSON, F.R.C.S., &c. *The Lancet*, December 13th, 1879, p. 873.

"HAD TO BE SUSPENDED."

JENNER had already observed phlegmonous erysipelas to follow vaccination. And in the Foundling Hospital at Petersburg, of 5,000 vaccinated infants, 57 took erysipelas in consequence of vaccination and died. Also in Boston, erysipelas has been seen to follow upon vaccination; and on various occasions vaccination has had to be suspended.—DR. C. SPINZU, St. Louis, U.S. 1886.

"AFRAID OF THE VIRUS."

One pernicious practice, is that of vaccinating children. I am as afraid of the virus as I am of the scallage. Not only scallage, but other impurities have been conveyed to the blood of healthy persons by means of vaccination. No one, who has given the subject the least consideration, doubts that impurities may be conveyed in this manner.—DR. A. G. SPRINGFIELD, Cleveland, Ohio. *New York Medical Tribune*, January, 1888.

"ITS GHASTLY RISKS."

There has fallen an ugly blot. It is too certain that one objection really formidable does exist—that the operation may, in some few instances, import to the subject of it the poison of a hateful and destructive disease (syphilis), peculiar to the human species, and the fear and nervous of its virus. The still distasteful subject I shall simply appeal to the printed testimony of MR. JONATHAN HUTCHINSON. . . . Such facts as he has demonstrated, con-

Q: What were the consequences of erysipelas following vaccination in the Foundling Hospital at Petersburg?

"HAD TO BE SUSPENDED."

JENNER had already observed phlegmonous erysipelas to follow vaccination. And in the Foundling Hospital at Petersburg, of 5,000 vaccinated infants, 57 took erysipelas in consequence of vaccination and died. Also in Boston, erysipelas has been seen to follow upon vaccination; and on various occasions vaccination has had to be suspended.—DR. C. SPINZU, St. Louis, U.S. 1886.

Evidence: (Page 3) And in the Foundling Hospital at Petersburg, of 5,000 vaccinated infants, 57 took erysipelas in consequence of vaccination and died.

A: 57 vaccinated infants died

GOT

- ✓ [OCR]: ... JENNER had already observed phlegmonous erysipelas to follow vaccination. And in the Foundling Hospital at Petersburg, of 5,000 vaccinated infants, 57 took erysipelas in consequence of vaccination and died. ...
- ✓ [R]: And in the Foundling Hospital at Petersburg, of 5,000 vaccinated infants, 57 took erysipelas in consequence of vaccination and died.
- ✓ [G]: Consequences: 57 infants died. Vaccination had to be suspended.

MinerU

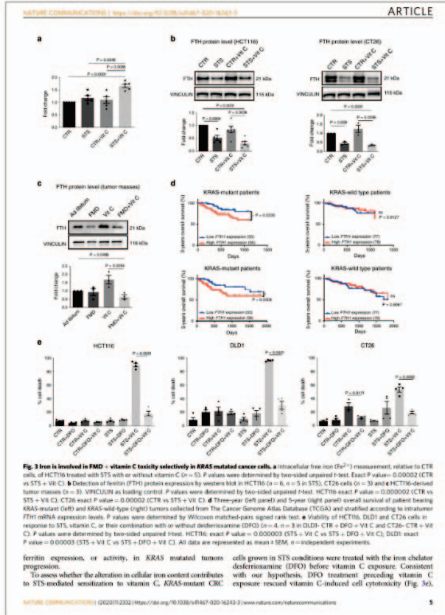
- ✓ [OCR]: # /n JENNER had already observed phlegmonous erysipelas to follow vaccination. And in the Foundling Hospital at Petersburg, of \$5,000\$ vaccinated infants, 57 took erysipelas in consequence of vaccination and died. ...
- ✓ [R]: of \$5,000\$ vaccinated infants, 57 took erysipelas in consequence of vaccination and died. ...
- ✓ [G]: 57 infants died.

Qwen2.5VL 72B

- ⚠ [OCR]: ...And in the Foundling Hospital at Petersburg, of 5,000 vaccinated infants, 57 took crysipelas in consequence of vaccination and died. Also in Boston,...
- ⚠ [R]: ... 57 took crysipelas in consequence of vaccination and died. Also in Boston...
- ✗ [G]: Consequences: Death, gangrenous condition, ulcerated sores, chronic eczema, debilitation, increased disease, fatal poisoning.

Figure S3. A case using text as the evidence source on a distorted academic document.

Evidence Source: Formula



Q: In the experiment assessing the viability of HCT116 cells treated with STS, vitamin C, and desferrioxamine (DFO), what is the exact P value for the comparison between STS + Vit C and STS + DFO + Vit C?

Fig. 3. HCTEs are involved in FMT + vitamin C protective selectively in KRAS mutated cancer cells. A: Intracellular free iron (Fe^{2+}) content, relative to CTR cells, of CTR16 was treated with FMT or with or without vitamin C ($n = 5$). P values were determined by two-sided unpaired *t*-test. Exact P value = 0.00002 (CTR vs FMT + VC). B: Detection of Heritin (Df) protein expression by western blot in HCT166 ($n = 5$), CTR16 ($n = 3$) and FMT + HCT16-derived CTR16 ($n = 3$). C: CTR16 vs FMT + CTR16 vs FMT + VC ($n = 3$). D: Three-year overall and 5-year (5-year split point) overall survival of patients bearing KRAS-mutated (left) and KRAS-wild type (right) tumors collected from The Cancer Genome Atlas Database (TCGA) and stratified according to intratumor Heritin mRNA expression levels. P values were determined by Wilcoxon matched-pairs signed rank test. * Viability of HCT16, DfD1 and CTR16 cells in the presence of FMT + VC or FMT + VC + Df. P values were determined by two-sided unpaired *t*-test. HCT16, exact P = 0.0000003 (FMT + VC vs FMT + VC + Df); exact P = 0.000003 (FMT + VC vs FMT + VC + Df + VC). All data are represented as mean \pm SEM, $n =$ independent experiments.

Evidence: (Page 4) HCT116: exact $\$P\$$ value $\$=0.000003\$$ (STS + Vit C vs STS + DFO + Vit C); DLD1: exact $\$P\$$ value $\$=0.000003\$$ (STS + Vit C vs STS + DFO + Vit C).

A: 0.000003.

GOT

✖ [OCR]: (Page2) \(\begin{array}{llllll}\text{HCT116} & \text{DLD1} & \text{CT26} & \text{b} & \text{SW48} & \text{HT29}\end{array}\backslash)\dots

[illegible]

✖ [G]: Not answerable.

MinerU F1-Score:1

✓ [OCR]:.. (\$n=4\$, \$n=3\$ in DLD1- CTR + DFO + Vit C and CT26- CTR + Vit C). P values were determined by two-sided unpaired t-test. HCT116: exact P value = 0.000003 (STS + Vit C vs STS + DFO + Vit C); DLD1: exact P value = 0.00003 (STS + Vit C vs STS + DFO + Vit C). All data are represented as mean \pm SEM, \$n=\$ independent experiments.

✓ [R]: ... HCT116: exact P value = 0.000003 (STS + Vit C vs STS + DFO + Vit C)...

✓ $[G]: 0.000003.$

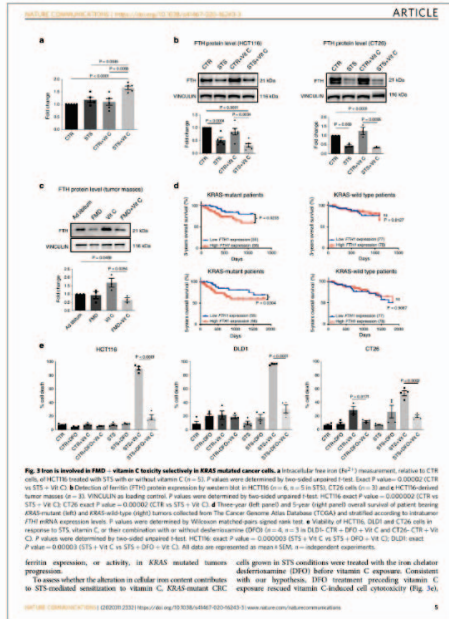
Qwen2.5VL 72B

✓ [OCR]: ... P values were determined by two-sided unpaired t-test. HCT116: exact P value = 0.000003 (STS + Vit C vs STS + DFO + Vit C); ...

✖ [R]: Wrong Retrieval Results. (Page 2) ...HCT116: exact \$P\$ value
\$=0.000000002\$ (Ad libitum vs FMD + Vit C)...

 [G]: 0.000000002.

Evidence Source: Formula



Q: In the experiment assessing the viability of HCT116 cells treated with STS, vitamin C, and desferrioxamine (DFO), what is the exact P value for the comparison between STS + Vit C and STS + DFO + Vit C?

Fig. 3. **3Tn** is involved in FMD-viral *in situ* toxicity selectively in KRAS mutated cancer cells. **a**, Intracellular free iron (Fe^{2+}) in the form of ferritin, relative to CTR cells and CTR cells treated with STS or without viral infection (C - vs S -) **P** values were determined by two-sided unpaired *t*-test. Exact *P* value = 0.000002 CTR vs STS - vs CTR - **b**, Detection of ferritin (CTR) protein expression by western blot in HCT116 (C - vs S -), STC76 cells (n = 3) and CTR cells (n = 3) and CTR cells-derived tumor masses (n = 3). VINCULIN as loading control. **P** values were determined by two-sided unpaired *t*-test. HCT116 exact *P* value = 0.0000002 CTR vs STS - vs CTR - STC76 exact *P* value = 0.0000002 CTR vs STS - vs CTR - **c**, KRAS-mutated (left) and KRAS-wild type (right) tumors collected from The Cancer Genome Atlas Database (TCGA) and stratified according to intratumoral Fe^{2+} mRNA expression levels. **P** values were determined by Wilcoxon matched-pairs signed rank test. **d**, Viability of HCT116, DLD1 and CTR6 cells in response to STS, V-ITC, or CTR, or their combination with STS + V-ITC, STS + CTR, V-ITC + CTR, STS + V-ITC + CTR. **e**, HCT116 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **f**, DLD1 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **g**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **h**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **i**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **j**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **k**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **l**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **m**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **n**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **o**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **p**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **q**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **r**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **s**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **t**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **u**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **v**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **w**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **x**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **y**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **z**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **aa**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ab**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ac**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ad**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ae**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **af**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ag**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ah**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ai**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **aj**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ak**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **al**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **am**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **an**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ao**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ap**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **aq**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ar**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **as**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **at**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **au**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **av**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **aw**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ax**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ay**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **az**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **ba**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **bb**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **bc**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **bd**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **be**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **bf**, CTR6 exact *P* value = 0.0000003 STS vs STS - CTR vs CTR - V-ITC vs V-ITC - STS + V-ITC vs STS + V-ITC - **bg**, CTR6 exact *P* value = 0.0000

Evidence: (Page 4) HCT116: exact $\$P\$$ value $\$=0.000003\$$ (STS + Vit C vs STS + DFO + Vit C); DLD1: exact $\$P\$$ value $\$=0.000003\$$ (STS + Vit C vs STS + DFO + Vit C).

A: 0.000003.

GOT

X [OCR]: (Page2) \(\begin{array}{lllll}\text{HCT116} & \& \text{DLD1} & \& \text{CT26} \\ & \& \text{b} & \& \text{SW48} & \& \text{HT29}\end{array}\backslash)\dots

[X]: (Page2) Fig. 1 FMD/STS enhances vitamin C anticancer activity in KRAS-mutant tumors. a Viability of KRAS-mutant and b KRAS-wild-type cancer cells treated for \((48\mathrm{~h})\) with STS....

✖ [G]: Not answerable.

MinerU

✓ [OCR]:.. (\$n=4\$, \$n=3\$ in DLD1- CTR + DFO + Vit C and CT26- CTR + Vit C). P values were determined by two-sided unpaired t-test. HCT116: exact P value = 0.000003 (STS + Vit C vs STS + DFO + Vit C); DLD1: exact P value = 0.00003 (STS + Vit C vs STS + DFO + Vit C). All data are represented as mean \pm SEM. \$n=\$ independent experiments.


✓ [R]: ... HCT116: exact P value = 0.000003 (STS + Vit C vs STS + DFO + Vit C)..

✓ **[G]: 0.000003.**

Qwen2.5VL 72B

✓ [OCR]: ... P values were determined by two-sided unpaired t-test. HCT116: exact P value = 0.000003 (STS + Vit C vs STS + DFO + Vit C); ...

✖ [R]: Wrong Retrieval Results. (Page 2) ...HCT116: exact \$P\$ value
\$=0.000000002\$ (Ad libitum vs FMD + Vit C)...

 [G]: 0.000000002.

Evidence Source: Text

NO.
Date:

2. "A或者B"类议论文模板:
 导入: 第一段: Some people hold the opinion that A is superior to B in many ways. Others, however, argue that B is much better. Personally, I would prefer A because I think A has more advantages.
 正文: 第二段: There are many reasons why I prefer A. The main reason is that... Another reason is that... (赞同A的原因)
 第三段: Of course, B also has advantage to some extent... (列举1~2个B的优势)
 结论: 第四段: But if all these factors are considered, A is much better than B. From what has been discussed above, we may finally draw the conclusion that... (得出结论)
 3. 观点论证类议论文模板:
 导入: 第一段: 提出一种现象或某个观点作为议论的话题
 As a student, I am seriously in favour of the decision (表明自己的观点是赞成还是反对)
 The reasons for this may be listed as follows (过渡句,承上启下)
 正文: 第二段: First of all... Secondly... Besides... (列举2~3个赞成或反对的理由)
 结论: 第三段: In conclusion, I believe that... (照应第一段,构成“总-分-总”结构)
 4. "How to"类议论文模板:
 导入: 第一段: 提出一种现象或某种困难作为议论的话题
 正文: 第二段: Many ways can help to solve this serious problem, but the following may be most effective. First of all... Another way to solve the problem is... Finally... (列举2~3个解决此问题的办法)

112

Q: In " 'A或者B' 类议论文模版", does the 'A or B' type argumentative essay template conclude by stating that A is better than B?

2. "A或者B"类议论文模板:
 导入: 第一段: Some people hold the opinion that A is superior to B in many ways. Others, however, argue that B is much better. Personally, I would prefer A because I think A has more advantages.
 正文: 第二段: There are many reasons why I prefer A. The main reason is that... Another reason is that... (赞同A的原因)
 第三段: Of course, B also has advantage to some extent... (列举1~2个B的优势)
 结论: 第四段: But if all these factors are considered, A is much better than B. From what has been discussed above, we may finally draw the conclusion that... (得出结论)

Evidence: (Page 0)结论: 第四段: But if all these factors are considered, A is much better than B. From what has been discussed above, we may finally draw the conclusion that: (得出结论)

A: Yes.

GOT

! [OCR]: "A或者B"类议论文模板: ... 结论: 第四段: But if all these factors are considered, A is much better than B. From what has been discussed above, we may finally draw the conclusion that. (得出结论)

! [R]: ... A is much better than B...

✓ [G]: Yes.

MinerU

✗ [OCR]: Can't Parse Handwritten Chinese.

✗ [R]: Wrong Retrieval Results.

✗ [G]: Not answerable

Qwen2.5VL 72B F1-Score:0

✓ [OCR]: "A 或者 B" 类议论文模板: ... 结论: 第四段: But if all these factors are considered, A is much better than B. From what has been discussed above, we may finally draw the conclusion that ... (得出结论)

✓ [R]: ...considered, A is much better than B...

✓ [G]: Yes.

Figure S7. A case using text as the evidence source on a handwritten textbook.

Evidence Source: Text

中考阅读理解各考管记	80
阅读训练的六种方法	80
1. 讲究阅读方法	80
2. 训练阅读速度	83
3. 改进阅读方式	84
4. 扩大英语阅读量	85
5. 强化阅读训练	85
6. 加强评估指导	86
做阅读理解的3点注意	86
"先题后文"阅读方法说明	87
阅读理解的答题技巧	87
如何应对阅读理解中的细节题和推理题	88
一. 做细节事实题的方法	88
二. 做推理判断题的方法	88
1. 事实推断	89
2. 替代推断	89
3. 逻辑推断	89
4. 对作者的态度和态度的推断	90
如何推测阅读理解题中的生词词义	90
一. 根据上下文或解释推测词义	90
二. 根据词量和逻辑进行判断	91
三. 根据并列或同位关系推测词义	91
四. 根据词量和常识判断	91
利用主题句解阅读理解题	92
一. 主题句在段落或篇首	92
二. 主题句在段落或篇末	92

Q: 在中考阅读理解备考中, 阅读训练的六种方法包括哪些?

中考阅读理解各考管记	80
阅读训练的六种方法	80
1. 讲究阅读方法	80
2. 训练阅读速度	83
3. 改进阅读方式	84
4. 扩大英语阅读量	85
5. 强化阅读训练	85
6. 加强评估指导	86

Evidence: (Page 0) 阅读训练的六种方法
80\n1. 讲究阅读方法 80\n2. 训练阅读速度 83\n3. 改进阅读方式 84\n4. 扩大英语阅读量 85\n5. 强化阅读训练 85\n6. 加强评估指导 86

A: 讲究阅读方法、训练阅读速度、改进阅读方式、扩大英语阅读量、强化阅读训练、加强评估指导

GOT

! [OCR]: 中青阅读理解各考管记\n\\begin{tabular}{|c|c|}\n\\hline 1. 讲究阅读方法 & 80\n\\hline 2. 训练阅读速度 & 83\n\\hline 3. 改进阅读方式 & 84\n\\hline 4. 扩大英语阅读量 & 85\n\\hline 5. 强化阅读训练 & 85\n\\hline 6. 加强评估指导 & \n\\hline\n\\end{tabular}

x [R] Wrong Retrieval Results.

x [G]: 阅读训练的六种方法包括: 通读全文、读题查读、简练作答、认真核查、答案形式符合提问方式、应对细节题和推理题的方法。

MinerU

x [OCR]: Can't Parse Handwritten Chinese.

x [R]: Wrong Retrieval Results.

x [G]: Not answerable.

Qwen2.5VL 72B

✓ [OCR]: \$\\begin{array}{|l|l|}\n\\hline\n\\text{中考阅读理解备考笔记} & 80\n\\hline\n\\text{阅读训练的六种方法} & 80\n\\hline\n\\text{1. 讲究阅读方法} & 80\n\\hline\n\\text{2. 训练阅读速度} & 83\n\\hline\n\\text{3. 改进阅读方式} & 84\n\\hline\n\\text{4. 扩大英语阅读量} & 85\n\\hline\n\\text{5. 强化阅读训练} & 85\n\\hline\n\\text{6. 加强评估指导} & \n\\hline\n\\end{array}

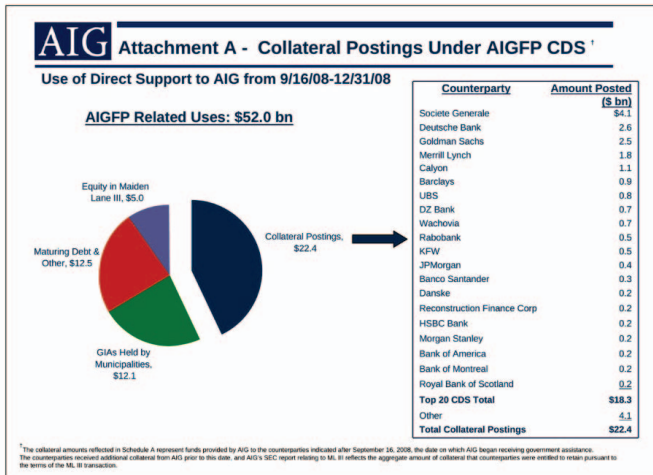
✓ [R]: \$\\begin{array}{|l|l|}\n\\hline\n\\text{中考阅读理解备考笔记} & 80\n\\hline\n\\text{阅读训练的六种方法} & 80\n\\hline\n\\text{1. 讲究阅读方法} & 80\n\\hline\n\\text{2. 训练阅读速度} & 83\n\\hline\n\\text{3. 改进阅读方式} & 84\n\\hline\n\\text{4. 扩大英语阅读量} & 85\n\\hline\n\\text{5. 强化阅读训练} & 85\n\\hline\n\\text{6. 加强评估指导} & \n\\hline\n\\end{array}

✓ [G]: \$训练阅读方法、训练阅读速度、改进阅读方式、扩大英语阅读量、强化阅读训练、加强评估指导。

Figure S8. A case using text as the evidence source on a handwritten textbook.

Evidence Source: Table

Q: What is the total dollar value of Collateral Postings under AIG's CDS?



Counterparty	Amount Posted (\$ bn)
Societe Generale	\$4.1
Deutsche Bank	2.6
Goldman Sachs	2.5
Merrill Lynch	1.8
Calyon	1.1
Bercleys	0.9
UBS	0.8
DZ Bank	0.7
Wachovia	0.7
Rabobank	0.5
KFW	0.5
JPMorgan	0.4
Banco Santander	0.3
Danske	0.2
Reconstruction Finance Corp	0.2
HSBC Bank	0.2
Morgan Stanley	0.2
Bank of America	0.2
Bank of Montreal	0.2
Royal Bank of Scotland	0.2
Top 20 CDS Total	\$18.3
Other	4.1
Total Collateral Postings	\$22.4

Evidence: (Page 2) Collateral Postings \$ 22.4

A: \$22.4 bn.

GOT

- ✓ [OCR]: \begin{tabular}{|c|c|} \hline Counterparty & Amount Posted \\ \hline Societe Generale & \\$4.1 \\ \hline Total Collateral Postings & \\$22.4 \\ \hline \end{tabular}
- ✓ [R]: ... \hline Other & 4.1 \\ \hline Total Collateral Postings & (\mathbf{\\$ 22.4}) \\ \hline
- ! [G]: Total Collateral Postings: \$22.4

MinerU

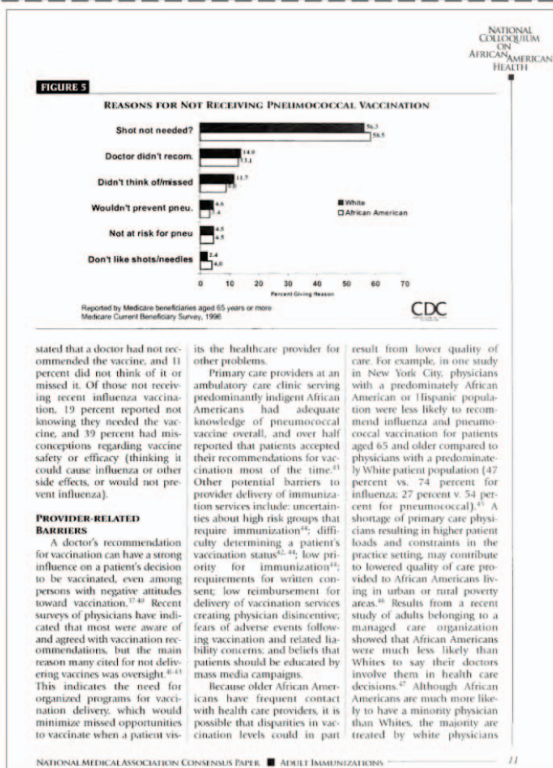
- ✗ [OCR]: Recognize the Table as Chart and Can't Parse Chart.
- ✗ [R]: Content in Other Document.
- ✗ [G]: Not answerable

Qwen2.5VL 72B

- ✓ [OCR]: \begin{tabular}{|c|c|} \hline Counterparty & Amount Posted \\ \hline Societe Generale & \$4.1 \\ \hline ... Total Collateral Postings & \$22.4 \\ \hline \end{tabular}
- ✓ [R]: Other & 4.1 \\ Total Collateral Postings & \$22.4 \\ \hline
- ✓ [G]: \$22.4 billion

Figure S9. A case using table as the evidence source on a financial report.

Evidence Source: Chart



Q: Among African American respondents, what reason was stated second most for not getting a pneumococcal vaccination?

Evidence: (Page 0)

```
<chart>\n\\begin{tabular}{| c |}\n\nReason & White & African\nAmerican \\ \\n\nShot not needed? & 56.3 & 58.5 \\ \\n\nDoctor didn't\nrecommend & 14.0 & 13.1 \\ \\n\nDidn't think of/missed & 11.7 & 9.0 \\ \\n\nWouldn't prevent pneumonia & 4.6 & 3.4 \\ \\n\nNot at risk for\npneumonia & 4.5 & 4.5 \\ \\n\nDon't like\nshots/needles & 2.4 & 4.0 \\ \\n\n\\end{tabular}</chart>"
```

A: Doctor didn't recommend

GOT

✗ [OCR]: **Incorrect Parsed Chart:** \section*{REASONS FOR NOT RECEIVING PNEUMOCOCCAL VACCINATION} \section*{Shot not needed?} Doctor didn't recom. \section*{Didn't think of missed} Wouldn't prevent pneu...

⚠ [R]: \section*{Shot not needed?} \nDoctor didn't recom. \n\n\section*{Didn't think of missed} \n\n\section*{Didn't think of missed}.

MinerU

✗ [OCR]: **Can't Parse Chart.**

✗ [R]: REASONS FOR NOT RECEIVING PNEUMOCOCCAL VACCINATION \n ...

✗ [G]: Not knowing they needed the vaccine.

Qwen2.5VL 72B

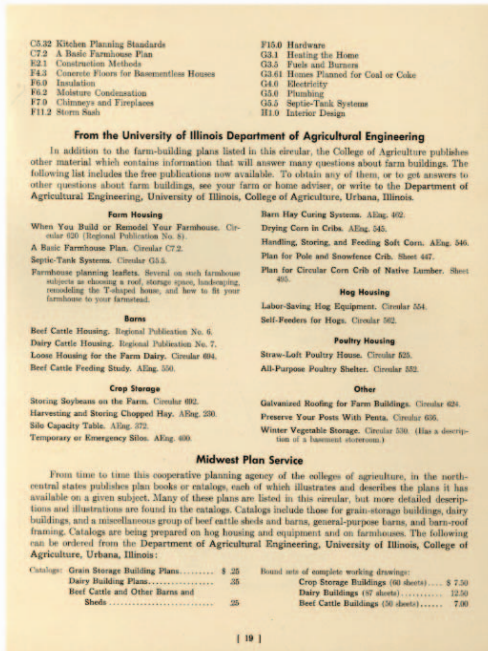
✓ [OCR]: ... \hline Shot not needed? & \multicolumn{2}{|c|}{56.3} \\ \hline Doctor didn't recom. & \multicolumn{2}{|c|}{14.0} \\ \hline Didn't think of/missed & \multicolumn{2}{|c|}{11.7} \\ \hline ...

✓ [R]: Doctor didn't recom. & \multicolumn{2}{|c|}{14.0} ...

✓ [G]: Doctor didn't recom.

Figure S10. A case using table as the evidence source on a scanned academic paper.

Evidence Source: Text



Q: What is the cost of obtaining the complete set of working drawings for Dairy Buildings from the University of Illinois Department of Agricultural Engineering?

Bound sets of complete working drawings:

Crop Storage Buildings (60 sheets)....	\$ 7.50
Dairy Buildings (87 sheets).....	12.50
Beef Cattle Buildings (50 sheets).....	7.00

Evidence(Page0): Bound sets of complete working drawings: Crop Storage Buildings (60 sheets) ... \\\$ 7.50 Dairy Buildings (87 sheets)..... 12.50 Beef Cattle Buildings (50 sheets) 7.00

A: \$12.50

GOT

- ✓ [OCR]: Crop Storage Buildings (60 sheets) \\\$ 7.50 Dairy Buildings (87 sheets)..... 12.50 Beef Cattle Buildings (50 sheets)..... 7.00.
- ✓ [R]: (page 18) ...Dairy Buildings (87 sheets)..... 12.50
- ✓ [G]: \$12.50

MinerU

- ✓ [OCR]: Bound sets of complete working drawings: Crop Storage Buildings (60 sheets).... \\\$s 7.50 Dairy Buildings (87 sheets)..... 12.50 Beef Cattle Buildings (50 sheets).... 7.00
- ✓ [R]: (page 18) ...Dairy Buildings (87 sheets)..... 12.50...
- ✓ [G]: The cost of obtaining the complete set of working drawings for Dairy Buildings from the University of Illinois Department of Agricultural Engineering is \$12.50.

Qwen2.5VL 72B

- ✓ [OCR]:Crop Storage Buildings (60 sheets).....\$ 7.50 \n Dairy Buildings (87 sheets).....12.50 \n Beef Cattle Buildings (50 sheets).....7.00
- ✓ [R]: (Page 18) Bound sets of complete working drawings:\n Crop Storage Buildings (60 sheets).....\$ 7.50\n Dairy Buildings (87 sheets).....12.50\nBeef Cattle Buildings (50 sheets).....7.00...
- ✗ [G]: Not answerable.

Figure S12. A case using text as the evidence source on a distortion manual.