Benchmarking Visual LLMs Resilience to Unanswerable Questions on Visually Rich Documents

Davide Napolitano, Luca Cagliero, Fabrizio Battiloro

Politecnico di Torino, Torino, Italy
name.surname@polito.it

1 Additional datasets' statistics

Table 1 reports the statistics of the original dataset and the selected subset. Table 2 reports the statistics for both the corrupted and verified datasets.

Tables 3, 4, 5 report additional statistics about the NLP entities, document elements, and layout information relative to each dataset.

Analysis of document element distribution (see Table 5) reveals a predominance of Abandon(headers, footers, footnotes, and marginal notes) and Text elements across both datasets, reflecting the underlying document types in the collections. Regarding entity (see Table 4), on both datasets, the most predominant ones are Numeric, Miscellaneous, and Location. The fine-grained entity distribution demonstrates both shared and distinct characteristics between the datasets. Familiar entities include measure units, person names, company names, spatial information, and document entity types. MPDocVQA shows higher frequencies of percentage-related entities, product references, and chemical elements, while DUDE exhibits a notable emphasis on means-of-transport-related entities. About layout characteristics (see Table 3), we observe an asymmetric distribution of entities, with a higher concentration in the left portions of documents. These distributional patterns persist consistently across both Corrupted and Verified versions of the datasets.

		MPDocVQA		DUDE	
		Full	**Sample**	**Full**	**Sample**
N° documents		5131	147	5017	277
N° pages	Avg	10.55	10.52	5.68	5.99
	Min	1	1	1	1
	Max	793	160	50	25
N° questions		36230	300	41453	300
N° questions / document	Avg	7.06	2.03	8.26	1.07
	Min	1	1	1	1
	Max	606	11	38	3

Table 1: Statistics about the original and sampled datasets

2 List of NLP entities

We analyze the effect of corrupting different NLP entities. To this end, we perform an extensive analysis of the sample datasets to identify prevalent topics and entity categories. Based on this analysis, we define a taxonomy of entities consisting of five categories:

- Numerical Corruption: "percentage", "currency", "temperature", "measure_unit", "numerical_value_number", "price_number_information", "price_numerical_value".
- Temporal Corruption: "date_information", "date_numerical_value", "time_information", "time_numerical_value", "year_number_information", "year_numerical_value"
- Entity Corruption: "person_name", "company_name", "product", "food", "chemical_element", "job_title_name", "job_title_information", "animal", "plant", "movie", "book", "transport_means", "event"
- Location Corruption: "country", "city", "street", "spatial_information", "continent", "postal_code_information", "postal_code_numerical_value"
- Document Structure Corruption: "document_position_information", "page_number_information", "page_number_numerical_value", "document_element_type", "document_element_information", "document_structure_information"

The implementation of the entity extraction phase based on GliNER (large v2) requires careful calibration of detection thresholds for specific entity types to optimize extraction quality. We establish entity-specific confidence thresholds with a default threshold of 0.75 for general entities. Document structure elements require a higher threshold (0.8) for "document_element_type", "document_element_information", and "document_structure_information". Similarly, for "postal_code_information" we set the threshold to 0.8, while for "postal_code_numerical_value" we set it to 0.78. For temporal entities, "date_information" we set the threshold to 0.75, while "year_numerical_value" we set it oto 0.7. Job-related entities required particularly stringent thresholds, i.e., for "job_title_name" 0.9, for "job_title_information" the threshold is 0.8, reflecting the complexity of accurately identifying these elements.

Given the absence of a comprehensive ground truth dataset for entity extraction in this context, we carry out a

Dataset	Version	Number of questions				Number of documents				Number of pages											
										Count			C1			C2			C3		
		Count	C1	C2	C3	Count	C1	C2	C3	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max
MPDocVQA	Corrupted	1408	840	434	134	82	82	65	25	5.95	1	40	6.00	1	40	5.65	1	21	6.22	1	21
	Verified	406	204	143	59	69	50	49	17	6.93	1	40	7.80	1	40	5.83	1	17	6.54	1	21
DUDE	Corrupted	768	495	199	74	87	87	44	15	5.33	1	20	5.45	1	20	4.85	1	17	5.89	1	10
	Verified	187	114	58	15	54	46	26	11	5.04	1	20	5.18	1	20	4.74	1	17	5.20	1	10

Table 2: Statistics about the corrupted and verified datasets. C*X* stands for Complexity=X.

	MPDocVQA						DUDE					
	Corrupted			Verified			Corrupted			Verified		
	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max
Top Left	13.02	0	89.00	13.44	0	70.00	10.56	0	49.00	10.14	0	36.00
Top Right	7.74	0	105.00	8.02	0	105.00	7.79	0	61.00	5.72	0	35.00
Bottom Left	10.90	0	104.00	11.05	0	59.00	10.08	0	49.00	9.74	0	38.00
Bottom Right	7.41	0	98.00	6.85	0	79.00	7.87	0	54.00	6.25	0	38.00

Table 3: Detailed layout information about the analyzed datasets.

manual evaluation and iterative refinement of both entity definitions and their associated detection thresholds. This process ensured high-quality entity extraction while maintaining the contextual relevance necessary for effective question corruption.

3 Experimental setup

In our experiment, we ensure maximal reproducibility and consistent evaluations across all models. For VLLMs, we standardize the token generation length to 1024 tokens to allow possible complete answers, while maintaining default settings for other parameters. The Qwen model implementation incorporated dynamic image scaling between 256 and 1440 pixels to optimize processing efficiency while preserving image quality. Llama 3.2 and Llava 1.6 are leveraged through the Ollama framework. To ensure comprehensive evaluation, each model is tested across all possible combinations of prompt configurations and window sizes. Concerning VLM, they are tested on the default setting, with a binary prompt and page-by-page. The binary prompting is forced to get that some corrupted questions are unanswerable, otherwise not possible.

Document Layout Analysis. Our document analysis pipeline employs DocLayout-YOLO for layout detection, configured with a deliberately low confidence threshold of 0.1 to maximize object detection coverage. This configuration ensures comprehensive capture of document elements, though it frequently results in overlapping detection boxes. To address this overlap, we implemented a refinement process that compares pairs of overlapping elements. When the intersection-over-union ratio exceeds 0.6, we retain the larger bounding box, ensuring optimal coverage while eliminating redundant detections.

OCR The text extraction process utilizes two specialized models based on content type. For standard textual elements, we employ GOT-OCR 2 with its OCR-specific configuration to ensure accurate text recognition. Visual elements, specifically figures and tables, undergo analysis using Qwen 2.5 VL 7B, configured with a 1024-token generation limit to produce detailed descriptive content. This dual-model approach ensures appropriate processing for both textual and visual document components while maintaining high-quality information extraction throughout the pipeline.

Table 4: NLP entity statistics over the datasets under analysis.

		MPDocVQA						DUDE					
		Corrupted			Verified			Corrupted			Verified		
		Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max
Macro Entities	**Numeric**	6.4	0	117.7	7.7	0	112.2	3.8	0	83.7	3.5	0	80.1
	Temporal	3.8	0	64.8	4.3	0	64.6	3.3	0	46.6	3.5	0	44.3
	Misc	9.9	0	175.1	10.5	0	128.5	7.7	0	154.0	6.0	0	61.6
	Location	6.4	0	99.5	6.7	0	72.0	8.8	0	129.1	7.5	0	65.0
	Structure	5.2	0	55.1	5.5	0	47.5	6.5	0	73.3	5.2	0	25.8
Numeric	number	5.0	0	137.0	5.3	0	137.0	3.0	0	57.0	2.7	0	39.0
	measure_unit	18.8	0	170.0	21.4	0	133.0	14.0	0	351.0	12.3	0	351.0
	price	0.9	0	33.0	1.2	0	33.0	0.6	0	14.0	0.5	0	10.0
	percentage	11.5	0	245.0	15.5	0	245.0	2.5	0	47.0	2.1	0	44.0
	temperature	0.9	0	15.0	0.9	0	14.0	1.2	0	25.0	1.0	0	25.0
	currency	7.5	0	224.0	9.8	0	224.0	5.5	0	92.0	5.6	0	92.0
Temporal	date	4.0	0	38.0	3.8	0	37.0	3.7	0	33.0	3.8	0	33.0
	time_info	8.4	0	104.0	8.7	0	104.0	8.3	0	105.0	9.9	0	105.0
	time_value	0.6	0	13.0	0.4	0	13.0	0.7	0	15.0	1.0	0	15.0
	year_info	1.5	0	47.0	1.6	0	47.0	1.0	0	21.0	0.8	0	7.0
	year_value	8.5	0	187.0	11.1	0	187.0	6.2	0	106.0	5.7	0	106.0
Miscellaneous	person	23.5	0	648.0	17.1	0	143.0	35.9	0	697.0	24.6	0	129.0
	company	24.7	0	347.0	26.6	0	347.0	14.1	0	112.0	11.7	0	63.0
	event	7.4	0	187.0	6.0	0	86.0	8.9	0	71.0	10.5	0	71.0
	product	13.9	0	109.0	17.1	0	109.0	6.7	0	273.0	3.0	0	42.0
	food	5.8	0	154.0	7.6	0	154.0	1.1	0	33.0	1.0	0	33.0
	chemical_elem	37.3	0	485.0	43.3	0	485.0	6.5	0	158.0	5.2	0	56.0
	job_title_name	5.7	0	104.0	6.2	0	104.0	6.2	0	61.0	6.2	0	39.0
	job_title_info	0.1	0	2.0	0.1	0	2.0	0.2	0	8.0	0.3	0	8.0
	animal	1.0	0	18.0	1.1	0	18.0	2.1	0	54.0	2.5	0	54.0
	plant	6.3	0	143.0	7.8	0	143.0	3.7	0	128.0	2.6	0	79.0
	movie	0.1	0	6.0	0.2	0	6.0	0.3	0	6.0	0.4	0	6.0
	book	1.3	0	25.0	1.4	0	25.0	3.3	0	190.0	1.0	0	9.0
	transport	2.3	0	49.0	2.1	0	49.0	11.1	0	212.0	9.0	0	212.0
Location	country	7.9	0	196.0	6.3	0	78.0	5.5	0	88.0	5.2	0	88.0
	city	7.7	0	137.0	7.2	0	62.0	7.0	0	68.0	6.6	0	63.0
	street	0.8	0	20.0	0.8	0	20.0	2.7	0	67.0	2.2	0	67.0
	spatial_info	22.1	0	163.0	24.3	0	163.0	43.2	0	609.0	35.8	0	201.0
	continent	4.4	0	153.0	5.4	0	153.0	2.0	0	30.0	1.9	0	27.0
	postal_code_info	2.1	0	26.0	2.5	0	26.0	1.4	0	41.0	0.7	0	9.0
	postal_code_val	0.0	0	2.0	0.0	0	2.0	0.0	0	1.0	0.0	0	0.0
Structure	doc_pos_info	4.6	0	64.0	4.4	0	34.0	4.6	0	50.0	3.2	0	28.0
	page_num_info	0.7	0	21.0	0.6	0	6.0	3.2	0	131.0	1.3	0	17.0
	page_num	0.0	0	1.0	0.0	0	0.0	0.0	0	6.0	0.0	0	0.0
	doc_elem_type	25.7	0	239.0	27.5	0	239.0	30.9	0	244.0	26.5	0	107.0
	doc_elem_info	0.3	0	4.0	0.3	0	4.0	0.4	0	9.0	0.2	0	3.0
	doc_struct_info	0.0	0	2.0	0.0	0	2.0	0.0	0	0.0	0.0	0	0.0

	MPDocVQA									DUDE								
	Augmented			Corrupted			Verified			Augmented			Corrupted			Verified		
	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max
abandon	16.59	0	218	17.74	0	218	19.52	0	218	10.50	0	75	8.22	0	36	7.09	0	36
figure	2.12	0	16	1.91	0	16	2.13	0	16	4.03	0	121	2.92	0	51	2.30	0	15
isolate_formula	0.10	0	3	0.12	0	3	0.09	0	3	0.17	0	6	0.10	0	4	0.13	0	4
plain text	27.50	0	312	29.29	0	312	28.49	0	213	31.18	0	285	29.52	0	192	25.59	0	121
table	1.52	0	38	1.81	0	38	2.06	0	38	1.37	0	19	1.47	0	19	1.26	0	13
title	5.89	0	64	6.68	0	64	7.23	0	64	8.19	0	97	6.14	0	32	5.54	0	25

Table 5: Document elements' statistics.

4 Prompt engineering

Corruption The corruption process occasionally produces syntactically or semantically challenged questions that require refinement to ensure human readability while maintaining their unanswerable nature. To address this challenge, we leverage the Qwen 2.5 7B language model. The model receives a carefully structured prompt that includes original and corrupted questions and explicit preservation instructions for corrupted elements. Our prompt engineering approach provides the model with several key components to ensure optimal refinement: (1) the original question for context, (2) the corrupted version requiring refinement, (3) a comprehensive list of corrupted elements that must remain unchanged, (4) specific refinement directives focusing on readability and natural language flow, and (5) carefully selected exemplars demonstrating both successful and unsuccessful refinements. This structured approach ensures that the refined questions maintain their intended unanswerable characteristics while achieving natural linguistic quality suitable for human evaluation.

```
PROMPT:
You are given two questions. The first one is the original one, the second one is the corrupted one.
The corruption is done based on entities extracted from the original question.

Original question: "{original_question}"
Corrupted question: "{corrupted_question}"

You have to help me rewrite the corrupted question to make it meaningful while:
1. Making it coherent and natural, while strictly keeping the exact same meaning
2. Ensuring it makes sense in the context of the original question
3. The following corrupted entities must be preserved in the rewritten question: {list(all_corrupted_entities)}
4. Editing the question minimally - only what's needed to make it coherent
5. Guaranteeing that the final output is meaningful

Original: "What is the highest temperature recorded?"
Bad corruption: "What is the 85 F temperature recorded?"
Correct rewrite: "Was 85 F the highest temperature recorded?"

Good Examples:
Original: "Which year is mentioned first in the x axis?"
Bad corruption: "Which 1975 is mentioned first in the x axis?"
Good rewrite: "Is 1975 the first year mentioned in the x axis?"

Original: "Which company had the most sales in 2022?"
Bad corruption: "Which Microsoft had the most sales in 2022?"
Correct rewrite: "Did Microsoft have the most sales in 2022?"

Important: Return only the rewritten question without any explanation or introductions.
```

Verification Our verification pipeline employs Gemini 2.5 Flash as an automated judge to evaluate the validity of corrupted questions. The verification process utilizes a structured prompt that incorporates several critical components to ensure accurate assessment. The prompt includes a detailed task description, comprehensive OCR output from the document page, and explicit entity mapping that shows the relationship between original and corrupted entities. To maintain spatial coherence during verification, we reconstruct the document's OCR content following the natural reading order, organizing text elements from top to bottom and left to right. This reconstruction approach is consistently applied across both the verification stage and subsequent VQA model evaluation, ensuring uniform document representation throughout the pipeline. The verification prompt specifies a standardized output format, facilitating automated processing of verification results while maintaining consistency across the evaluation pipeline. This structured approach ensures reliable identification by looking at "verifi-

cation_result" field, set to false if the corrupted question is unanswerable.

```
PROMPT:
You are an expert in Visual Question Answering on Document images.
We are working on a project to verify the answerability of questions based on the information provided in a given image.
In detail we have taken questions from a multipage VQA dataset and we have corrupted the questions based on the entities found in the whole document associated to the question.
Now, given the corrupted question and each image of the document, we want to verify if the question is answerable based solely on the information provided in the given image.
Your task is to help us to determine if the following corrupted question is answerable based solely on the information provided in the given image.
The question answer must be explicitly stated in the image.
In order to have a better document understanding, we extracted the following OCR text from the document :\n{ocr_text}

In addition here we provide the original entities found in the question and the corrupted ones in order to allow you to place special focus on the corrupted ones. The entities are reported with the format: ORIGINAL -- $>$ CORRUPTED:\n{entities_string}

Respond with a structured response in JSON format with the following fields :
{
    "verification_result": "true if the question is answerable based solely on the information provided in the given image, or 'false' if it's not answerable",
    "question_answer": "The answer to the question or only the words 'not found' if the answer is not explicitly stated in the image"
}
Return only the JSON response. Without any other text or explanation.
Question: {question}
```

Questions marked as unanswerable were manually validated by three NLP experts (MSc or higher), achieving 96.97% precision

VLLM For Vision-Language Large Language Models (VLLMs), we implemented a comprehensive evaluation framework that systematically tests different prompt configurations within defined context windows. Our experimental design explores the impact of two key factors: explicit notification of potential question unanswerability and the inclusion of document OCR text. The base prompt template establishes a clear task context and role definition for the model while maintaining flexibility for our experimental conditions:

```
PROMPT:
You are an AI assistant specialized in analyzing document images and text.
Your task is to answer questions about the document image content precisely.

For this question, you have the following OCR text: {ocr_text} # OPTIONAL

Guidelines:
- Provide concise, focused answers (single word or short phrase preferred)
- Base your answer on both the image and the provided OCR text
- If uncertain, return 'Unable to determine' # OPTIONAL
- If you can't find the answer, return 'Unable to determine' # OPTIONAL
Question: {question}
```

This template incorporates several key elements: task specification, role definition, optional OCR context, and structured response guidelines. The optional components allow for a systematic evaluation of how different context levels affect model performance. To ensure optimal performance while maintaining comparability, we adapted the base prompt structure according to each model's author-recommended prompting patterns, while preserving the core evaluation framework.

Ouput Standardization To process metrics, we need a standard output. Although properly prompted, VLLMs may not follow output format directives. To overcome this issue, we leverage an LLM-as-a-judge that standardizes outputs that are not properly formatted. This is done by exploiting Gemini 2.0 Flash with the following prompt:

```
PROMPT:
I'm performing an evaluation test on the ability of different models to answer VQA questions from document images.
The model could return different answers to determine if the answer is 'unable to determine' or not.
Your task is to detect if the answer means that the model is unable to determine the answer or not.
Examples of answers that mean that the model is unable to determine the answer:
- Not available.
- Not provided in document.
- The image does not provide information to answer the question.
```

```
- I cannot provide an answer based on the given text.
- The document does not provide information
If the answer means 'unable to determine', respond with 'unable to determine', otherwise return the original answer.
The answer is: {answer}
Please respond only with the original answer or 'unable to determine' only.
```

5 Additional results

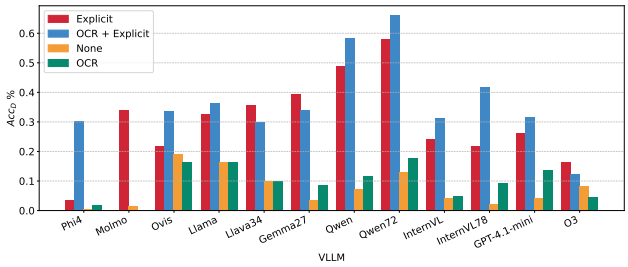
RQ2 - Document and Page-Level Accuracy Table 6 and 7 provide fine details about performances on analyzed metrics, respectively at document and page level. In detail, they extend the radar plots reported in the main paper by adding VLM performance. As expected, they perform poorly due to their nature and task settings.

RQ2/RQ3 - Document-Level Ablation In Table 8 and 10, we report the ablation study on the different models for different prompts and complexity levels. To reduce the cumbersome quantity of data and focus on relevant results, we decide to place focus on the two prompt types where the unanswertability is made explicit since providing the most relevant results (see Research Question 3 in the main paper).

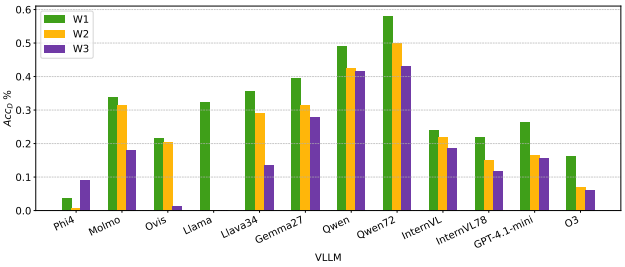
The reported results demonstrate a clear performance advantage for Qwen when augmented with OCR explicit information, consistently achieving superior document-level accuracy across varying complexity conditions. This suggests that the integration of explicit text recognition significantly enhances document comprehension capabilities beyond what can be achieved through visual processing alone. Performance degradation is evident as document complexity increases from C1 to C3, though this effect varies across models. The substantial gap between OCR-enhanced and standard approaches underscores the importance of text recognition in document understanding tasks. Models exhibit heterogeneous performance patterns based on document characteristics, with notable sensitivities to document length, where accuracy typically diminishes as page count increases beyond 8 pages. Entity-based analysis reveals differential performance across semantic categories. Location entities are generally processed more effectively, while Structure entities present consistent challenges across most models. This pattern manifests similarly in both datasets, suggesting fundamental limitations in how current vision-language models process structural document information. Interestingly, documents with lower element density ($<$15%) yield better performance, indicating that visual clutter adversely affects comprehension capabilities. The comparative analysis between DUDE and MPDocVQA demonstrates that while general performance trends remain consistent, the latter dataset shows less pronounced degradation across complexity levels for certain models, suggesting dataset-specific characteristics influence model robustness.

RQ2/RQ3 - Page-Level Ablation The ablation studies on page-level accuracy across DUDE and MPDocVQA datasets (Table 9, 11) demonstrate consistent superiority of Qwen with OCR explicit integration, highlighting the transformative impact of combining visual processing with textual recognition. This performance advantage persists across varying complexity levels, though it becomes less pronounced at C3, where models like Llava and Gemma sometimes outperform Qwen, suggesting these models possess enhanced resilience to extreme complexity. The integration of OCR capabilities produces asymmetric benefits across document characteristics. For instance, while providing substantial improvements for most models on text-heavy elements, its impact on figures and tables is less consistent. This pattern indicates fundamental differences in how models process textual versus visual information in documents, with OCR integration primarily enhancing text extraction capabilities rather than comprehensive visual understanding. Document element density emerges as a significant performance determinant, with most models achieving superior results on documents with lower element density ($<$15%). This finding suggests that visual clutter presents a substantial challenge for current vision-language models. The spatial positioning of information also significantly impacts performance, with bottom-right positions generally yielding better results, potentially due to reading pattern biases in model training data. Entity type analysis reveals pronounced performance differentials, with Numeric and Temporal entities being processed effectively while Structure entities remain challenging. This disparity indicates that current architectures excel at extracting discrete information but struggle with understanding document organization and hierarchical relationships. Notably, the MPDocVQA dataset shows less pronounced performance degradation across complexity levels compared to DUDE, suggesting dataset-specific characteristics influence model robustness. In-page analysis further demonstrates that document understanding is highly context-dependent, with models exhibiting different strengths based on element type and position.

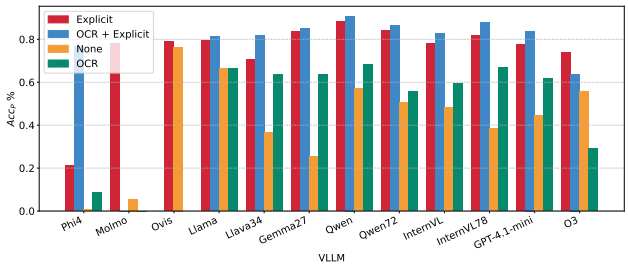
RQ2/RQ3 - In-Page Ablation The in-page analyses on Table 12, 13 reveal that document understanding is highly element-dependent and spatially nuanced, with consistent patterns emerging across both datasets despite their distinct characteristics. Element-type analysis demonstrates that contemporary models exhibit specialized processing capabilities for different document components. Title elements generally yield the highest accuracy in DUDE, likely due to their distinctive visual formatting and semantic importance, while tables present persistent challenges that suggest limitations in structural reasoning. Interestingly, MPDocVQA shows strong table recognition capabilities for several models, indicating dataset-specific training or representation factors influence element processing capabilities. Spatial positioning emerges as a critical factor in document understanding, with elements positioned in the bottom-right quadrant consistently achieving higher accuracy across models and complexity levels. This phenomenon reflects the same correlation between document elements and layout observed in the main paper. OCR integration provides substantial but non-uniform benefits across elements and po-



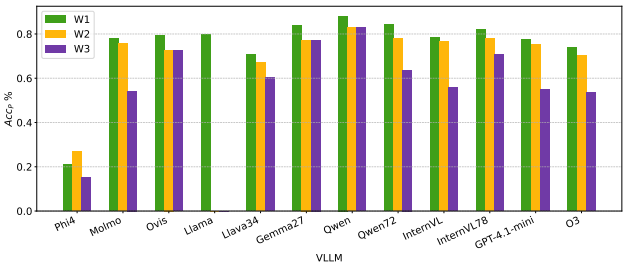
(a) MPDocVQA - Document Level Accuracy - Ablation parameters



(b) MPDocVQA - Document Level Accuracy - Ablation windows



(c) MPDocVQA - Page Level Accuracy - Ablation parameters



(d) MPDocVQA - Page Level Accuracy - Ablation windows

Figure 1: Ablation study on in-context learning strategy and window size. MPDocVQA dataset (addressing RQ3)

sitions. Text-heavy elements show the most consistent improvements with OCR, while the benefits for figures are less pronounced. This differential impact highlights the complementary nature of visual and textual processing in document understanding tasks. The integration appears more consistently beneficial in MPDocVQA compared to DUDE, suggesting dataset characteristics influence the utility of explicit text recognition. Complexity resilience varies significantly across element types and spatial positions. While performance generally degrades from C1 to C3, certain elements and positions maintain robust accuracy even at higher complexity levels. MPDocVQA demonstrates superior complexity resilience compared to DUDE, particularly for abandoned elements and bottom-positioned content. This difference suggests that dataset design characteristics substantially impact model robustness to document complexity.

These findings collectively underscore the multifaceted nature of document understanding, revealing that current vision-language models process documents through a complex interplay of element recognition, spatial reasoning, and textual integration. Future architectural improvements should focus on enhancing structural understanding capabilities and mitigating spatial biases to advance fine-grained document comprehension performance.

RQ3 - Ablation study on MPDocVQA Our study on in-context learning for vision-language models reveals striking patterns in unanswerable question detection (Figure 1). Explicitly stating that questions may be unanswerable dramatically improves model performance. Including OCR-extracted text substantially boosts accuracy across all conditions, providing critical context for answerability determination. Combining explicit unanswerability instructions with OCR integration produces the strongest results, revealing powerful synergy between task understanding and information access. Counterintuitively, page-level accuracy plummets as window size increases—suggesting current models struggle when larger contexts dilute essential information.

References

		Phi4	Molmo	Ovis	Llama	Llava 34B	Gemma3 27B	Qwen2.5 VL 7B	Qwen2.5 VL 72B	InterVL3 9B	InterVL3 78B	GPT4.1 mini	O3
DUDE													
	Acc_D	0.070	0.230	0.241	0.289	0.401	0.503	0.460	0.599	0.267	0.326	0.214	0.239
Doc El	<15%	0.032	0.168	0.248	0.272	0.408	0.512	0.384	0.592	0.216	0.312	0.192	0.177
	15%-25%	0.000	0.048	0.008	0.048	0.072	0.040	0.096	0.080	0.048	0.032	0.016	0.073
	>25%	0.072	0.128	0.104	0.112	0.120	0.200	0.208	0.224	0.136	0.144	0.112	0.045
Layout	<4 pages	0.080	0.243	0.309	0.273	0.416	0.556	0.493	0.624	0.240	0.340	0.210	0.122
	4-8 pages	0.019	0.101	0.178	0.197	0.267	0.310	0.317	0.368	0.202	0.248	0.108	0.116
	>8 pages	0.000	0.069	0.000	0.069	0.232	0.417	0.347	0.458	0.200	0.042	0.014	0.039
NLP Entity	Numeric	0.001	0.500	0.143	0.357	0.286	0.357	0.714	0.929	0.357	0.286	0.143	0.185
	Temporal	0.064	0.170	0.170	0.191	0.553	0.511	0.426	0.638	0.191	0.362	0.340	0.248
	Misc	0.120	0.270	0.390	0.340	0.460	0.610	0.580	0.790	0.350	0.410	0.300	0.312
	Location	0.020	0.204	0.204	0.306	0.469	0.735	0.429	0.510	0.306	0.367	0.122	0.196
	Structure	0.031	0.123	0.031	0.123	0.185	0.185	0.200	0.215	0.108	0.092	0.062	0.003
MPDocVQA													
		Phi4	Molmo	Ovis	Llama	Llava 34B	Gemma3 27B	Qwen2.5 VL 7B	Qwen2.5 VL 72B	InterVL3 9B	InterVL3 78B	GPT4.1 mini	O3
	Acc_D	0.037	0.340	0.217	0.325	0.357	0.394	0.490	0.581	0.241	0.219	0.264	0.163
Doc El	<15%	0.033	0.354	0.227	0.331	0.340	0.392	0.492	0.572	0.243	0.213	0.265	0.166
	15%-25%	0.000	0.019	0.019	0.028	0.094	0.075	0.056	0.104	0.019	0.028	0.009	0.019
	>25%	0.037	0.100	0.050	0.112	0.149	0.124	0.187	0.224	0.100	0.112	0.124	0.050
Layout	<4 pages	0.042	0.296	0.218	0.316	0.567	0.514	0.500	0.655	0.252	0.232	0.234	0.176
	4-8 pages	0.059	0.402	0.360	0.414	0.159	0.556	0.569	0.598	0.331	0.331	0.468	0.355
	>8 pages	0.041	0.440	0.141	0.370	0.131	0.235	0.526	0.571	0.286	0.200	0.317	0.166
NLP Entity	Numeric	0.007	0.340	0.163	0.340	0.313	0.299	0.442	0.565	0.143	0.197	0.197	0.116
	Temporal	0.149	0.511	0.277	0.553	0.340	0.383	0.638	0.660	0.362	0.255	0.468	0.298
	Misc	0.019	0.256	0.207	0.298	0.369	0.343	0.421	0.515	0.184	0.146	0.201	0.117
	Location	0.038	0.454	0.308	0.346	0.431	0.608	0.685	0.692	0.400	0.300	0.338	0.215
	Structure	0.118	0.265	0.176	0.265	0.412	0.265	0.235	0.529	0.176	0.147	0.206	0.088

Table 6: Effect of the corruption type on the Document-Level Accuracy. Coarse-grained analysis (addressing RQ2).

Table 7: Effect of the corruption type on the Page-Level Accuracy. Fine-grained analysis (addressing RQ2).

		DUDE											
		Phi4	Molmo	Ovis	Llama	Llava 34B	Gemma3 27B	Qwen2.5 VL 7B	Qwen2.5 VL 72B	InterVL3 9B	InterVL3 78B	GPT4.1 mini	O3
	Acc_P	0.248	0.554	0.674	0.680	0.717	0.786	0.835	0.754	0.713	0.781	0.638	0.663
Doc El	0	0.247	0.532	0.746	0.692	0.755	0.813	0.867	0.777	0.773	0.850	0.753	0.709
	1	0.240	0.555	0.627	0.658	0.685	0.784	0.812	0.759	0.661	0.692	0.531	0.575
	>1	0.263	0.614	0.550	0.684	0.667	0.713	0.784	0.737	0.632	0.737	0.497	0.577
Lay	In-Page	0.207	0.444	0.566	0.536	0.655	0.740	0.753	0.802	0.579	0.701	0.500	0.522
	Out-Page	0.267	0.606	0.725	0.748	0.747	0.808	0.873	0.700	0.777	0.819	0.703	0.712
NLP Entity	Numeric	0.236	0.906	0.890	0.866	0.661	0.906	0.969	0.990	0.906	0.906	0.638	0.662
	Temporal	0.299	0.528	0.492	0.563	0.787	0.650	0.787	0.736	0.614	0.711	0.690	0.652
	Misc	0.233	0.448	0.724	0.678	0.724	0.856	0.848	0.858	0.701	0.767	0.626	0.529
	Location	0.183	0.409	0.668	0.545	0.800	0.902	0.851	0.713	0.749	0.770	0.574	0.484
	Structure	0.233	0.571	0.568	0.682	0.673	0.652	0.774	0.602	0.647	0.741	0.641	0.727
		MPDocVQA											
		Phi	Qwen	Molmo	InternVL	DocOwl	Ovis	Llama	Gemma	Llava	UDOP	LayoutLMv3	BLIP
	Acc_P	0.211	0.780	0.792	0.796	0.708	0.838	0.881	0.842	0.782	0.818	0.775	0.738
Doc El	0	0.231	0.761	0.839	0.772	0.726	0.869	0.881	0.851	0.808	0.864	0.793	0.761
	1	0.203	0.794	0.769	0.823	0.700	0.807	0.889	0.858	0.776	0.784	0.758	0.714
	>1	0.154	0.817	0.667	0.812	0.658	0.803	0.858	0.832	0.699	0.736	0.751	0.716
Lay	In-Page	0.184	0.620	0.638	0.638	0.705	0.758	0.800	0.792	0.609	0.661	0.577	0.563
	Out-Page	0.221	0.835	0.844	0.850	0.709	0.865	0.909	0.878	0.842	0.872	0.842	0.798
NLP Entity	Numeric	0.258	0.820	0.799	0.829	0.715	0.836	0.890	0.842	0.766	0.810	0.766	0.773
	Temporal	0.276	0.944	0.897	0.949	0.774	0.850	0.970	0.950	0.909	0.899	0.937	0.923
	Misc	0.161	0.668	0.776	0.702	0.657	0.829	0.829	0.813	0.702	0.792	0.713	0.675
	Location	0.182	0.809	0.703	0.752	0.749	0.825	0.904	0.819	0.797	0.775	0.682	0.603
	Structure	0.258	0.682	0.732	0.778	0.783	0.768	0.843	0.801	0.758	0.793	0.773	0.692

			Phi4		Molmo	Ovis		Llama		Llava 34B		Gemma 27B		Qwen 2.5 7B		Qwen 2.5 72B		InternVL 3 9B		InternVL 3 78B		GPT-4.1-mini		O3	
			Explicit	OCR Explicit	Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit
	Acc_D	C1	0.079	0.439	0.254	0.281	0.509	0.342	0.325	0.377	0.325	0.482	0.430	0.465	0.570	0.588	0.649	0.281	0.404	0.342	0.404	0.202	0.298	0.227	0.319
		C2	0.052	0.534	0.190	0.172	0.190	0.224	0.362	0.483	0.431	0.586	0.310	0.517	0.672	0.707	0.741	0.259	0.397	0.328	0.397	0.276	0.276	0.301	0.080
		C3	0.067	0.133	0.200	0.200	0.133	0.133	0.267	0.267	0.267	0.333	0.133	0.200	0.267	0.267	0.467	0.200	0.200	0.200	0.133	0.067	0.067	0.092	0.066
Document Element	<15%	C1	0.042	0.306	0.208	0.306	0.500	0.333	0.278	0.389	0.278	0.514	0.319	0.389	0.431	0.583	0.611	0.236	0.347	0.333	0.347	0.167	0.278	0.319	0.297
		C2	0.023	0.419	0.093	0.186	0.186	0.209	0.302	0.442	0.395	0.535	0.233	0.442	0.605	0.698	0.698	0.186	0.395	0.302	0.395	0.279	0.279	0.256	0.353
		C3	0.023	0.419	0.093	0.186	0.186	0.209	0.302	0.442	0.395	0.535	0.233	0.442	0.605	0.698	0.698	0.186	0.395	0.302	0.395	0.279	0.279	0.256	0.296
	15%-25%	C1	0.000	0.040	0.040	0.010	0.110	0.040	0.020	0.080	0.010	0.020	0.080	0.100	0.130	0.080	0.120	0.040	0.070	0.030	0.090	0.010	0.010	0.168	0.002
		C2	0.000	0.080	0.160	0.000	0.080	0.160	0.080	0.080	0.080	0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.080	0.160	0.080	0.080	0.157	0.050
		C3	0.000	0.080	0.160	0.000	0.080	0.160	0.080	0.080	0.080	0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.080	0.160	0.080	0.080	0.033	0.068
	>25%	C1	0.078	0.310	0.129	0.116	0.142	0.142	0.194	0.090	0.207	0.207	0.233	0.194	0.271	0.220	0.233	0.142	0.181	0.155	0.155	0.129	0.168	0.183	0.096
		C2	0.052	0.310	0.129	0.052	0.052	0.052	0.181	0.207	0.181	0.233	0.155	0.233	0.284	0.233	0.284	0.129	0.103	0.129	0.103	0.078	0.078	0.031	0.031
		C3	0.052	0.310	0.129	0.052	0.052	0.052	0.181	0.207	0.181	0.233	0.155	0.233	0.284	0.233	0.284	0.129	0.103	0.129	0.103	0.078	0.078	0.031	0.183
Layout	<4 pages	C1	0.099	0.626	0.273	0.424	0.308	0.363	0.319	0.446	0.403	0.525	0.516	0.606	0.581	0.643	0.836	0.282	0.667	0.373	0.542	0.210	0.280	0.234	0.315
		C2	0.036	0.446	0.179	0.107	0.143	0.143	0.304	0.429	0.357	0.536	0.268	0.393	0.536	0.607	0.607	0.179	0.196	0.250	0.250	0.196	0.179	0.005	0.002
		C3	0.062	0.125	0.125	0.188	0.125	0.125	0.250	0.375	0.438	0.312	0.125	0.188	0.312	0.250	0.562	0.125	0.188	0.188	0.125	0.062	0.062	0.110	0.095
	4-8 pages	C1	0.000	0.258	0.113	0.148	0.547	0.220	0.258	0.330	0.258	0.261	0.148	0.294	0.398	0.368	0.401	0.187	0.223	0.223	0.148	0.038	0.181	0.018	0.143
		C2	0.056	0.333	0.000	0.222	0.111	0.167	0.222	0.111	0.222	0.389	0.111	0.444	0.611	0.444	0.333	0.278	0.889	0.278	0.556	0.222	0.278	0.121	0.210
		C3	0.000	0.000	0.167	0.000	0.000	0.000	0.167	0.167	0.000	0.167	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.031	0.061
	>8 pages	C1	0.000	0.103	0.026	0.000	0.709	0.026	0.000	0.167	0.051	0.359	0.251	0.344	0.410	0.462	0.436	0.118	0.318	0.077	0.251	0.000	0.077	0.091	0.180
		C2	0.000	0.111	0.333	0.000	0.333	0.333	0.000	0.111	0.111	0.370	0.333	0.407	0.481	0.481	0.556	0.370	0.370	0.000	0.407	0.037	0.000	0.104	0.093
		C3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.333	0.000	0.333	0.000	0.000	0.000	0.000	0.167	0.333	0.000	0.000	0.000	0.000	0.000	0.147	0.084
NLP Entity	Numeric	C1	0.000	0.000	0.250	0.250	0.750	0.250	0.250	0.125	0.250	0.125	0.500	0.750	0.875	1.000	1.000	0.250	0.500	0.250	0.500	0.125	0.125	0.102	0.094
		C2	0.000	0.500	0.833	0.000	0.333	0.500	0.500	0.500	0.333	0.667	0.667	0.667	0.833	0.833	1.000	0.500	0.500	0.333	0.500	0.167	0.167	0.173	0.200
		C3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.142	0.098
	Temporal	C1	0.091	0.545	0.364	0.182	0.455	0.364	0.364	0.545	0.545	0.636	0.636	0.364	0.545	0.727	0.909	0.273	0.364	0.455	0.364	0.182	0.182	0.127	0.134
		C2	0.042	0.583	0.083	0.167	0.333	0.167	0.333	0.708	0.667	0.583	0.250	0.583	0.667	0.792	0.750	0.208	0.125	0.417	0.500	0.542	0.250	0.553	0.231
		C3	0.083	0.167	0.167	0.167	0.083	0.083	0.417	0.250	0.167	0.250	0.083	0.167	0.167	0.250	0.667	0.083	0.167	0.167	0.083	0.083	0.083	0.032	0.027
	Misc	C1	0.139	0.583	0.361	0.528	0.556	0.444	0.528	0.528	0.556	0.667	0.528	0.556	0.722	0.861	0.917	0.389	0.500	0.472	0.556	0.389	0.611	0.426	0.491
		C2	0.098	0.569	0.196	0.275	0.196	0.275	0.392	0.451	0.412	0.667	0.373	0.627	0.745	0.784	0.784	0.333	0.529	0.353	0.451	0.275	0.431	0.214	0.513
		C3	0.154	0.308	0.308	0.462	0.308	0.308	0.308	0.308	0.692	0.231	0.308	0.462	0.462	0.615	0.538	0.308	0.462	0.462	0.308	0.154	0.154	0.125	0.289
	Location	C1	0.036	0.536	0.179	0.250	0.750	0.393	0.214	0.393	0.143	0.679	0.464	0.500	0.643	0.429	0.464	0.321	0.429	0.357	0.393	0.143	0.250	0.083	0.243
		C2	0.000	0.750	0.250	0.125	0.125	0.188	0.438	0.562	0.438	0.812	0.375	0.375	0.750	0.750	0.812	0.125	0.250	0.438	0.188	0.125	0.188	0.310	0.288
		C3	0.000	0.000	0.200	0.200	0.200	0.200	0.200	0.600	0.200	0.800	0.200	0.200	0.200	0.200	0.400	0.800	0.200	0.200	0.200	0.000	0.000	0.001	0.036
	Structure	C1	0.065	0.258	0.161	0.065	0.194	0.194	0.194	0.194	0.161	0.129	0.194	0.290	0.258	0.258	0.323	0.129	0.258	0.161	0.226	0.065	0.065	0.041	0.000
		C2	0.000	0.211	0.053	0.000	0.000	0.105	0.211	0.211	0.211	0.158	0.053	0.211	0.368	0.316	0.474	0.158	0.474	0.053	0.263	0.105	0.000	0.188	0.036
		C3	0.000	0.000	0.133	0.000	0.000	0.000	0.133	0.133	0.000	0.333	0.000	0.000	0.200	0.000	0.267	0.000	0.000	0.000	0.000	0.000	0.000	0.122	0.154

Table 8: Effect of the corruption type on the Document-Level Accuracy by varying in-context learning strategy and complexity level (addressing RQ2 and RQ3). DUDE dataset.

			Phi4		Molmo	Ovis		Llama		Llava 34B		Gemma 27B		Qwen 2.5 7B		Qwen 2.5 72B		InternVL 3 9B		InternVL 3 78B		GPT-4.1-mini		O3	
			Explicit	OCR Explicit	Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit
	Acc_P	C1	0.266	0.750	0.577	0.723	0.638	0.712	0.699	0.701	0.738	0.810	0.772	0.843	0.870	0.753	0.812	0.738	0.783	0.805	0.827	0.636	0.724	0.661	0.618
		C2	0.240	0.818	0.542	0.615	0.593	0.655	0.742	0.760	0.771	0.764	0.724	0.847	0.920	0.816	0.896	0.684	0.778	0.760	0.800	0.669	0.691	0.694	0.773
		C3	0.141	0.692	0.423	0.513	0.551	0.526	0.654	0.692	0.744	0.679	0.692	0.731	0.731	0.519	0.712	0.628	0.603	0.667	0.667	0.538	0.590	0.563	0.468
Document Element	0	C1	0.244	0.674	0.498	0.787	0.581	0.694	0.735	0.725	0.773	0.842	0.780	0.859	0.873	0.688	0.682	0.766	0.784	0.856	0.863	0.742	0.808	0.895	0.828
		C2	0.283	0.799	0.610	0.667	0.623	0.698	0.767	0.799	0.830	0.748	0.736	0.887	0.962	0.958	0.958	0.767	0.824	0.836	0.887	0.767	0.811	0.744	0.615
		C3	0.283	0.799	0.610	0.667	0.623	0.698	0.767	0.799	0.830	0.748	0.736	0.887	0.962	0.958	0.958	0.767	0.824	0.836	0.887	0.767	0.811	0.744	0.679
	1	C1	0.264	0.862	0.621	0.707	0.667	0.741	0.672	0.678	0.730	0.787	0.822	0.833	0.868	0.719	0.842	0.736	0.816	0.759	0.816	0.557	0.667	0.715	0.686
		C2	0.198	0.840	0.469	0.556	0.556	0.556	0.691	0.704	0.716	0.827	0.691	0.815	0.877	0.881	0.881	0.593	0.741	0.605	0.691	0.519	0.543	0.595	0.617
		C3	0.198	0.840	0.469	0.556	0.556	0.556	0.691	0.704	0.716	0.827	0.691	0.815	0.877	0.881	0.881	0.593	0.741	0.605	0.691	0.519	0.543	0.472	0.560
	>1	C1	0.317	0.770	0.698	0.595	0.730	0.714	0.651	0.675	0.667	0.770	0.683	0.817	0.865	0.804	0.889	0.675	0.738	0.754	0.762	0.500	0.611	0.554	0.600
		C2	0.143	0.857	0.400	0.514	0.543	0.686	0.743	0.714	0.629	0.686	0.743	0.743	0.829	0.667	0.851	0.514	0.657	0.771	0.657	0.571	0.486	0.525	0.456
		C3	0.143	0.857	0.400	0.514	0.543	0.686	0.743	0.714	0.629	0.686	0.743	0.743	0.829	0.667	0.851	0.514	0.657	0.771	0.657	0.571	0.486	0.525	0.338
Layout	In-Page	C1	0.254	0.725	0.551	0.696	0.551	0.645	0.681	0.638	0.681	0.761	0.768	0.775	0.833	0.816	0.832	0.623	0.739	0.746	0.775	0.478	0.645	0.502	0.573
		C2	0.185	0.758	0.371	0.492	0.500	0.492	0.621	0.677	0.718	0.790	0.653	0.782	0.911	0.869	0.897	0.548	0.653	0.694	0.710	0.565	0.573	0.373	0.527
		C3	0.119	0.595	0.310	0.357	0.405	0.310	0.524	0.643	0.643	0.524	0.643	0.595	0.643	0.533	0.756	0.524	0.476	0.571	0.571	0.381	0.452	0.208	0.558
	Out-Page	C1	0.269	0.757	0.585	0.731	0.664	0.733	0.704	0.720	0.755	0.826	0.773	0.863	0.881	0.712	0.798	0.773	0.797	0.823	0.843	0.684	0.748	0.628	0.783
		C2	0.285	0.868	0.682	0.715	0.669	0.788	0.841	0.828	0.815	0.742	0.781	0.901	0.927	0.679	0.893	0.795	0.881	0.815	0.874	0.755	0.788	0.654	0.612
		C3	0.167	0.806	0.556	0.694	0.722	0.778	0.806	0.750	0.861	0.861	0.750	0.889	0.833	0.429	0.429	0.750	0.750	0.778	0.778	0.722	0.750	0.754	0.782
NLP Entity	Numeric	C1	0.235	0.864	0.864	0.901	0.975	0.840	0.605	0.605	0.753	0.877	0.951	0.975	0.988	1.000	1.000	0.901	0.951	0.901	0.951	0.617	0.765	0.526	0.727
		C2	0.239	0.935	0.978	0.870	0.870	0.913	0.848	0.761	0.826	0.957	0.957	0.957	0.978	0.974	1.000	0.913	0.935	0.913	0.913	0.674	0.783	0.533	0.715
		C3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.147	0.061
	Temporal	C1	0.273	0.727	0.394	0.394	0.606	0.636	0.606	0.788	0.636	0.758	0.606	0.697	0.667	0.885	0.962	0.545	0.576	0.606	0.697	0.455	0.394	0.432	0.497
		C2	0.359	0.803	0.556	0.453	0.479	0.504	0.658	0.855	0.863	0.641	0.573	0.855	0.932	0.800	0.859	0.607	0.641	0.718	0.778	0.744	0.658	0.750	0.751
		C3	0.170	0.702	0.553	0.660	0.702	0.660	0.723	0.617	0.660	0.596	0.681	0.681	0.681	0.414	0.828	0.681	0.660	0.766	0.745	0.723	0.702	0.581	0.618
	Misc	C1	0.310	0.814	0.540	0.823	0.743	0.735	0.805	0.805	0.823	0.876	0.814	0.858	0.912	0.940	0.952	0.743	0.814	0.832	0.841	0.699	0.858	0.645	0.827
		C2	0.184	0.821	0.400	0.732	0.658	0.689	0.774	0.711	0.716	0.900	0.811	0.874	0.932	0.856	0.906	0.700	0.837	0.758	0.784	0.653	0.726	0.664	0.759
		C3	0.244	0.644	0.422	0.444	0.467	0.489	0.556	0.578	0.756	0.622	0.622	0.711	0.644	0.600	0.560	0.600	0.600	0.644	0.600	0.333	0.489	0.218	0.586
	Location	C1	0.214	0.717	0.434	0.711	0.887	0.610	0.635	0.811	0.723	0.912	0.836	0.893	0.912	0.693	0.781	0.767	0.792	0.824	0.818	0.635	0.730	0.673	0.682
		C2	0.222	0.861	0.389	0.444	0.500	0.556	0.667	0.667	0.667	0.917	0.667	0.639	0.833	0.750	0.812	0.417	0.583	0.472	0.500	0.194	0.389	0.134	0.370
		C3	0.025	0.800	0.325	0.700	0.725	0.275	0.600	0.875	0.900	0.850	0.900	0.875	0.900	0.767	0.800	0.975	0.825	0.825	0.850	0.675	0.800	0.646	0.689
	Structure	C1	0.293	0.698	0.624	0.659	0.259	0.741	0.741	0.580	0.712	0.678	0.654	0.766	0.800	0.571	0.654	0.678	0.727	0.771	0.800	0.639	0.683	0.579	0.563
		C2	0.224	0.783	0.609	0.559	0.540	0.671	0.752	0.770	0.776	0.602	0.677	0.826	0.901	0.742	0.887	0.714	0.807	0.814	0.870	0.739	0.714	0.924	0.796
		C3	0.127	0.667	0.402	0.402	0.451	0.578	0.686	0.706	0.716	0.676	0.647	0.706	0.725	0.431	0.681	0.480	0.490	0.569	0.588	0.490	0.500	0.489	0.636

Table 9: Effect of the corruption type on the Page-Level Accuracy by varying in-context learning strategy and complexity level (addressing RQ2). DUDE dataset.

			Phi4		Molmo		Ovis		Llama		Llava 34B		Gemma 27B		Qwen 2.5 7B		Qwen 2.5 72B		InternVL 3 9B		InternVL 3 78B		GPT-4.1-mini		O3	
			Explicit	OCR Explicit	Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit
	Acc_D	C1	0.044	0.348	0.358	0.221	0.451	0.314	0.387	0.309	0.289	0.402	0.377	0.500	0.598	0.613	0.632	0.255	0.328	0.275	0.402	0.294	0.363	0.186	0.137
		C2	0.028	0.259	0.329	0.189	0.224	0.322	0.350	0.441	0.329	0.420	0.301	0.497	0.573	0.538	0.685	0.259	0.336	0.175	0.434	0.259	0.280	0.168	0.126
		C3	0.034	0.254	0.305	0.271	0.220	0.373	0.322	0.322	0.271	0.305	0.305	0.441	0.559	0.576	0.695	0.153	0.203	0.136	0.441	0.169	0.254	0.068	0.068
Document Element	<15%	C1	0.033	0.328	0.367	0.228	0.444	0.317	0.394	0.278	0.267	0.383	0.344	0.494	0.594	0.600	0.622	0.244	0.322	0.261	0.394	0.289	0.356	0.178	0.133
		C2	0.031	0.248	0.341	0.194	0.240	0.318	0.349	0.450	0.326	0.442	0.279	0.519	0.566	0.550	0.690	0.271	0.349	0.171	0.426	0.271	0.287	0.186	0.132
		C3	0.031	0.248	0.341	0.194	0.240	0.318	0.349	0.450	0.326	0.442	0.279	0.519	0.566	0.550	0.690	0.271	0.349	0.171	0.426	0.271	0.287	0.186	0.132
	15%-25%	C1	0.000	0.100	0.020	0.020	0.040	0.040	0.060	0.120	0.080	0.120	0.100	0.040	0.080	0.120	0.140	0.040	0.040	0.060	0.040	0.020	0.040	0.040	0.020
		C2	0.000	0.064	0.032	0.032	0.000	0.032	0.032	0.064	0.032	0.000	0.064	0.000	0.064	0.032	0.128	0.000	0.000	0.000	0.064	0.000	0.032	0.000	0.000
		C3	0.000	0.064	0.032	0.032	0.000	0.032	0.032	0.064	0.032	0.000	0.064	0.000	0.064	0.032	0.128	0.000	0.000	0.000	0.064	0.000	0.032	0.000	0.000
	>25%	C1	0.063	0.147	0.126	0.063	0.210	0.105	0.105	0.147	0.147	0.147	0.210	0.231	0.231	0.231	0.210	0.126	0.147	0.126	0.189	0.147	0.168	0.084	0.063
		C2	0.000	0.112	0.075	0.037	0.037	0.149	0.149	0.112	0.149	0.112	0.187	0.149	0.261	0.187	0.187	0.075	0.112	0.112	0.187	0.075	0.075	0.000	0.037
		C3	0.000	0.112	0.075	0.037	0.037	0.149	0.149	0.112	0.149	0.112	0.187	0.149	0.261	0.187	0.187	0.075	0.112	0.112	0.187	0.075	0.075	0.000	0.037
Layout	<4 pages	C1	0.029	0.487	0.327	0.279	0.344	0.344	0.472	0.527	0.345	0.619	0.498	0.550	0.654	0.735	0.796	0.268	0.371	0.359	0.482	0.309	0.373	0.223	0.149
		C2	0.058	0.215	0.321	0.156	0.152	0.302	0.390	0.654	0.271	0.450	0.307	0.477	0.592	0.578	0.783	0.280	0.338	0.091	0.419	0.227	0.232	0.181	0.091
		C3	0.073	0.227	0.199	0.136	0.172	0.267	0.168	0.321	0.235	0.266	0.266	0.266	0.420	0.430	0.537	0.176	0.108	0.181	0.387	0.023	0.126	0.023	0.000
	4-8 pages	C1	0.125	0.250	0.250	0.312	0.250	0.350	0.287	0.188	0.250	0.512	0.350	0.450	0.550	0.613	0.512	0.250	0.350	0.250	0.550	0.450	0.613	0.350	0.250
		C2	0.000	0.381	0.508	0.381	0.381	0.452	0.397	0.127	0.381	0.579	0.508	0.651	0.579	0.579	0.635	0.381	0.381	0.381	0.437	0.468	0.437	0.341	0.286
		C3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	>8 pages	C1	0.047	0.298	0.456	0.151	0.616	0.331	0.264	0.149	0.281	0.200	0.327	0.506	0.569	0.466	0.503	0.295	0.316	0.211	0.355	0.303	0.358	0.143	0.205
		C2	0.018	0.179	0.196	0.018	0.161	0.167	0.232	0.036	0.196	0.277	0.196	0.310	0.420	0.375	0.482	0.161	0.196	0.179	0.342	0.161	0.241	0.018	0.036
		C3	0.000	0.094	0.219	0.062	0.104	0.302	0.208	0.125	0.000	0.000	0.198	0.292	0.333	0.375	0.333	0.094	0.125	0.000	0.135	0.125	0.167	0.094	0.125
NLP Entity	Numeric	C1	0.000	0.216	0.324	0.054	0.432	0.189	0.351	0.189	0.189	0.216	0.216	0.486	0.622	0.676	0.622	0.108	0.189	0.270	0.405	0.162	0.135	0.054	0.000
		C2	0.000	0.339	0.339	0.210	0.258	0.339	0.387	0.403	0.339	0.371	0.226	0.435	0.452	0.500	0.597	0.161	0.274	0.210	0.355	0.210	0.258	0.177	0.129
		C3	0.021	0.250	0.354	0.188	0.208	0.458	0.354	0.292	0.271	0.271	0.354	0.417	0.500	0.562	0.792	0.146	0.208	0.125	0.500	0.208	0.312	0.083	0.083
	Temporal	C1	0.130	0.435	0.609	0.304	0.696	0.565	0.565	0.304	0.522	0.304	0.435	0.652	0.870	0.696	0.739	0.391	0.435	0.304	0.609	0.522	0.565	0.391	0.478
		C2	0.267	0.533	0.467	0.200	0.400	0.600	0.533	0.400	0.667	0.467	0.467	0.533	0.667	0.600	0.800	0.200	0.333	0.267	0.667	0.400	0.333	0.133	0.200
		C3	0.000	0.556	0.333	0.333	0.333	0.444	0.556	0.333	0.556	0.444	0.444	0.778	0.889	0.667	0.889	0.556	0.667	0.111	0.667	0.444	0.667	0.333	0.444
	Misc	C1	0.023	0.322	0.299	0.253	0.506	0.333	0.402	0.345	0.241	0.402	0.356	0.471	0.529	0.598	0.609	0.230	0.322	0.230	0.345	0.276	0.402	0.195	0.126
		C2	0.007	0.125	0.236	0.139	0.132	0.257	0.292	0.410	0.222	0.354	0.236	0.451	0.514	0.479	0.653	0.201	0.257	0.104	0.340	0.208	0.243	0.125	0.069
		C3	0.038	0.205	0.244	0.282	0.192	0.333	0.282	0.321	0.282	0.256	0.333	0.308	0.462	0.487	0.590	0.103	0.141	0.128	0.397	0.103	0.192	0.013	0.000
	Location	C1	0.093	0.465	0.442	0.302	0.349	0.326	0.349	0.256	0.395	0.605	0.558	0.628	0.698	0.628	0.651	0.419	0.465	0.395	0.465	0.372	0.442	0.233	0.140
		C2	0.000	0.407	0.556	0.296	0.407	0.426	0.389	0.574	0.481	0.685	0.519	0.722	0.833	0.704	0.815	0.593	0.648	0.315	0.704	0.444	0.426	0.315	0.241
		C3	0.030	0.212	0.303	0.333	0.242	0.242	0.212	0.424	0.182	0.485	0.152	0.697	0.758	0.758	0.758	0.061	0.121	0.152	0.364	0.121	0.121	0.030	0.000
	Structure	C1	0.000	0.357	0.143	0.071	0.071	0.071	0.214	0.571	0.143	0.429	0.286	0.071	0.214	0.357	0.571	0.071	0.143	0.143	0.214	0.143	0.143	0.000	0.000
		C2	0.273	0.455	0.182	0.182	0.091	0.182	0.455	0.455	0.455	0.182	0.273	0.273	0.636	0.636	0.818	0.000	0.182	0.091	0.455	0.091	0.091	0.000	0.182
		C3	0.111	0.556	0.556	0.333	0.333	0.667	0.667	0.111	0.222	0.111	0.222	0.444	0.667	0.667	0.667	0.556	0.556	0.222	0.556	0.444	0.556	0.333	0.444

Table 10: Effect of the corruption type on the Document-Level Accuracy by varying in-context learning strategy and complexity level (addressing RQ2 and RQ3). MPDocVQA dataset.

			Phi4		Molmo		Ovis		Llama		Llava 34B		Gemma 27B		Qwen 2.5 7B		Qwen 2.5 72B		InternVL 3 9B		InternVL 3 78B		GPT-4.1-mini		O3	
			Explicit	OCR Explicit	Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit
	Acc_P	C1	0.225	0.807	0.830	0.815	0.682	0.845	0.857	0.707	0.849	0.852	0.883	0.901	0.925	0.855	0.864	0.829	0.869	0.849	0.897	0.827	0.871	0.780	0.687
		C2	0.188	0.706	0.699	0.740	0.759	0.724	0.743	0.729	0.763	0.824	0.807	0.850	0.887	0.791	0.834	0.725	0.778	0.757	0.843	0.691	0.776	0.669	0.558
		C3	0.205	0.728	0.749	0.808	0.813	0.749	0.795	0.663	0.798	0.808	0.824	0.865	0.889	0.885	0.918	0.712	0.751	0.824	0.891	0.741	0.842	0.712	0.611
Document Element	0	C1	0.247	0.714	0.815	0.858	0.614	0.829	0.855	0.733	0.837	0.875	0.869	0.889	0.918	0.837	0.848	0.854	0.879	0.875	0.916	0.829	0.882	0.797	0.698
		C2	0.215	0.667	0.709	0.809	0.822	0.728	0.787	0.733	0.804	0.870	0.841	0.883	0.922	0.844	0.925	0.802	0.848	0.846	0.915	0.767	0.863	0.743	0.626
		C3	0.215	0.667	0.709	0.809	0.822	0.728	0.787	0.733	0.804	0.870	0.841	0.883	0.922	0.844	0.925	0.802	0.848	0.846	0.915	0.767	0.863	0.743	0.626
	1	C1	0.232	0.897	0.850	0.826	0.703	0.876	0.868	0.698	0.879	0.847	0.915	0.932	0.944	0.863	0.863	0.844	0.888	0.861	0.900	0.842	0.888	0.777	0.685
		C2	0.146	0.729	0.649	0.625	0.653	0.688	0.653	0.715	0.681	0.747	0.743	0.795	0.830	0.870	0.861	0.604	0.663	0.604	0.733	0.562	0.635	0.549	0.444
		C3	0.146	0.729	0.649	0.625	0.653	0.688	0.653	0.715	0.681	0.747	0.743	0.795	0.830	0.870	0.861	0.604	0.663	0.604	0.733	0.562	0.635	0.549	0.444
	>1	C1	0.131	0.841	0.822	0.636	0.841	0.804	0.832	0.650	0.799	0.794	0.827	0.850	0.888	0.860	0.870	0.701	0.776	0.729	0.822	0.771	0.785	0.734	0.659
		C2	0.186	0.837	0.814	0.756	0.779	0.826	0.814	0.756	0.814	0.837	0.837	0.860	0.895	0.724	0.762	0.721	0.791	0.791	0.826	0.709	0.779	0.674	0.570
		C3	0.186	0.837	0.814	0.756	0.779	0.826	0.814	0.756	0.814	0.837	0.837	0.860	0.895	0.724	0.762	0.721	0.791	0.791	0.826	0.709	0.779	0.674	0.570
Layout	In-Page	C1	0.223	0.750	0.695	0.719	0.512	0.734	0.781	0.719	0.730	0.836	0.820	0.855	0.898	0.863	0.878	0.695	0.770	0.754	0.832	0.703	0.746	0.664	0.473
		C2	0.142	0.606	0.582	0.551	0.609	0.591	0.618	0.717	0.612	0.698	0.668	0.766	0.791	0.692	0.764	0.566	0.625	0.563	0.708	0.495	0.625	0.511	0.425
		C3	0.210	0.609	0.572	0.696	0.710	0.572	0.681	0.652	0.645	0.754	0.739	0.775	0.819	0.836	0.895	0.551	0.587	0.717	0.826	0.536	0.717	0.500	0.406
	Out-Page	C1	0.225	0.818	0.856	0.833	0.714	0.866	0.872	0.705	0.872	0.856	0.894	0.910	0.930	0.852	0.858	0.855	0.888	0.868	0.909	0.850	0.895	0.802	0.728
		C2	0.218	0.770	0.774	0.861	0.855	0.809	0.823	0.737	0.859	0.904	0.896	0.904	0.949	0.929	0.933	0.827	0.876	0.880	0.929	0.815	0.872	0.770	0.642
		C3	0.202	0.794	0.847	0.871	0.871	0.847	0.859	0.669	0.883	0.839	0.871	0.915	0.927	0.947	0.947	0.802	0.843	0.883	0.927	0.855	0.911	0.831	0.726
NLP Entity	Numeric	C1	0.277	0.802	0.845	0.808	0.652	0.854	0.878	0.710	0.838	0.857	0.878	0.912	0.942	0.877	0.873	0.811	0.860	0.866	0.905	0.799	0.838	0.784	0.631
		C2	0.258	0.753	0.771	0.760	0.784	0.773	0.776	0.714	0.742	0.836	0.789	0.854	0.857	0.754	0.789	0.695	0.740	0.745	0.794	0.703	0.745	0.732	0.612
		C3	0.239	0.799	0.855	0.836	0.843	0.871	0.874	0.720	0.855	0.814	0.877	0.912	0.925	0.904	0.956	0.805	0.821	0.830	0.912	0.808	0.865	0.811	0.714
	Temporal	C1	0.240	0.957	0.974	0.908	0.962	0.974	0.951	0.770	0.962	0.862	0.954	0.980	0.992	0.957	0.974	0.921	0.941	0.918	0.957	0.951	0.959	0.949	0.934
		C2	0.611	0.903	0.806	0.819	0.875	0.889	0.889	0.833	0.931	0.861	0.889	0.903	0.931	0.868	0.925	0.778	0.861	0.750	0.931	0.847	0.861	0.750	0.708
		C3	0.183	0.963	0.927	0.908	0.899	0.899	0.954	0.752	0.963	0.798	0.954	0.982	0.991	0.968	0.989	0.954	0.954	0.927	0.963	0.945	0.972	0.945	0.927
	Misc	C1	0.175	0.728	0.761	0.813	0.655	0.788	0.827	0.633	0.800	0.841	0.843	0.859	0.889	0.809	0.809	0.790	0.846	0.822	0.876	0.799	0.860	0.760	0.661
		C2	0.137	0.633	0.638	0.758	0.761	0.693	0.718	0.688	0.734	0.818	0.808	0.818	0.873	0.785	0.832	0.705	0.767	0.758	0.831	0.688	0.781	0.665	0.535
		C3	0.186	0.631	0.618	0.765	0.780	0.628	0.719	0.629	0.756	0.834	0.788	0.814	0.842	0.861	0.891	0.603	0.659	0.818	0.877	0.665	0.829	0.601	0.475
	Location	C1	0.226	0.770	0.840	0.741	0.486	0.819	0.794	0.733	0.827	0.864	0.901	0.909	0.926	0.813	0.825	0.852	0.868	0.823	0.881	0.782	0.856	0.613	0.490
		C2	0.151	0.832	0.817	0.670	0.724	0.735	0.742	0.846	0.839	0.828	0.828	0.943	0.964	0.849	0.881	0.857	0.885	0.789	0.939	0.659	0.824	0.627	0.556
		C3	0.158	0.526	0.705	0.705	0.674	0.632	0.600	0.505	0.526	0.716	0.589	0.779	0.832	0.750	0.781	0.484	0.589	0.611	0.747	0.495	0.568	0.505	0.337
	Structure	C1	0.297	0.766	0.453	0.578	0.094	0.609	0.688	0.875	0.734	0.828	0.750	0.719	0.734	0.444	0.556	0.625	0.672	0.688	0.734	0.625	0.656	0.547	0.453
		C2	0.171	0.686	0.314	0.429	0.486	0.571	0.743	0.800	0.771	0.714	0.629	0.771	0.886	0.789	0.895	0.429	0.457	0.600	0.743	0.543	0.429	0.257	0.257
		C3	0.263	0.960	0.960	0.939	0.939	0.960	0.970	0.717	0.919	0.747	0.929	0.949	0.970	0.963	0.963	0.960	0.960	0.929	0.960	0.949	0.960	0.939	0.939

Table 11: Effect of the corruption type on the Page-Level Accuracy by varying in-context learning strategy and complexity level (addressing RQ2). MPDocVQA dataset.

			Phi4		Molmo		Ovis		Llama		Llava 34B		Gemma 27B		Qwen 2.5 7B		Qwen 2.5 72B		InternVL 3 9B		InternVL 3 78B		GPT-4.1-mini		O3	
			Explicit	OCR Explicit	Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit
Document Element	Title	C1	0.500	1.000	0.750	1.000	0.500	0.750	1.000	0.750	0.750	0.750	0.750	1.000	1.000	1.000	1.000	0.500	0.750	1.000	1.000	0.500	0.750	0.550	0.788
		C2	0.000	0.750	0.125	0.500	0.500	0.375	0.625	0.500	0.625	0.750	0.625	0.750	0.875	0.917	0.917	0.375	0.500	0.750	0.625	0.250	0.250	0.236	0.190
		C3	0.000	1.000	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	1.000	0.500	0.500	0.500	1.000	0.500	0.500	0.500	1.000	0.500	1.000	0.565	0.971
	Text	C1	0.149	0.649	0.486	0.730	0.662	0.608	0.662	0.689	0.622	0.811	0.730	0.757	0.824	0.804	0.813	0.581	0.770	0.743	0.797	0.419	0.635	0.571	0.575
		C2	0.179	0.731	0.358	0.478	0.522	0.478	0.567	0.746	0.746	0.746	0.567	0.776	0.910	0.880	0.890	0.507	0.612	0.687	0.687	0.582	0.567	0.559	0.752
		C3	0.059	0.588	0.294	0.529	0.588	0.176	0.471	0.706	0.706	0.647	0.765	0.765	0.706	0.783	0.826	0.824	0.588	0.824	0.765	0.529	0.706	0.506	0.705
	Figure	C1	0.464	0.857	0.607	0.679	0.536	0.714	0.750	0.571	0.857	0.679	0.714	0.750	0.893	0.944	0.917	0.643	0.643	0.714	0.750	0.500	0.679	0.658	0.573
		C2	0.296	0.815	0.333	0.333	0.259	0.481	0.667	0.667	0.630	1.000	0.630	0.815	0.852	0.792	0.917	0.667	0.593	0.519	0.704	0.296	0.519	0.373	0.601
		C3	0.000	0.778	0.000	0.000	0.111	0.111	0.556	0.667	0.778	0.444	0.778	0.556	0.667	0.545	0.818	0.667	0.667	0.667	0.556	0.000	0.444	0.047	0.322
	Table	C1	0.250	0.800	0.650	0.450	0.400	0.450	0.500	0.450	0.400	0.650	0.850	0.750	0.750	0.667	0.792	0.550	0.700	0.700	0.650	0.400	0.400	0.454	0.421
		C2	0.000	0.810	0.381	0.429	0.476	0.524	0.619	0.667	0.857	0.714	0.762	0.810	0.952	1.000	1.000	0.476	0.619	0.762	0.714	0.810	0.667	0.763	0.471
		C3	0.211	0.526	0.421	0.316	0.316	0.474	0.632	0.579	0.579	0.474	0.526	0.474	0.579	0.333	0.714	0.211	0.316	0.368	0.421	0.368	0.263	0.322	0.130
	Abandon	C1	0.286	0.714	0.571	0.857	0.357	0.786	0.786	0.643	0.857	0.786	0.857	0.857	0.857	0.625	0.625	0.857	0.857	0.857	0.857	0.714	0.857	0.738	0.877
		C2	0.333	0.556	0.667	1.000	0.889	0.778	0.667	0.333	0.444	0.889	0.889	0.778	1.000	0.800	0.800	0.667	1.000	0.889	0.889	0.667	0.778	0.475	0.852
		C3	0.000	0.429	0.429	0.286	0.286	0.286	0.571	0.571	0.571	0.429	0.571	0.429	0.571	0.556	0.889	0.429	0.571	0.286	0.429	0.429	0.571	0.256	0.589
Layout	Top Left	C1	0.184	0.632	0.658	0.737	0.500	0.711	0.553	0.658	0.526	0.868	0.816	0.763	0.842	0.660	0.680	0.658	0.763	0.763	0.763	0.526	0.658	0.470	0.646
		C2	0.053	0.687	0.321	0.237	0.305	0.382	0.481	0.687	0.718	0.763	0.481	0.611	0.870	0.970	0.970	0.260	0.382	0.603	0.664	0.748	0.580	0.647	0.550
		C3	0.053	0.687	0.321	0.237	0.305	0.382	0.481	0.687	0.718	0.763	0.481	0.611	0.870	0.970	0.970	0.260	0.382	0.603	0.664	0.748	0.580	0.780	0.432
	Top Right	C1	0.280	0.760	0.440	0.720	0.560	0.640	0.640	0.600	0.760	0.840	0.760	0.800	0.800	0.857	0.886	0.600	0.720	0.760	0.640	0.480	0.560	0.389	0.488
		C2	0.038	0.846	0.423	0.462	0.538	0.538	0.654	0.769	0.885	0.923	0.808	0.962	1.000	0.967	0.967	0.577	0.462	0.692	0.615	0.538	0.538	0.397	0.492
		C3	0.038	0.846	0.423	0.462	0.538	0.538	0.654	0.769	0.885	0.923	0.808	0.962	1.000	0.967	0.967	0.577	0.462	0.692	0.615	0.538	0.538	0.685	0.644
	Bottom Left	C1	0.180	0.689	0.541	0.639	0.393	0.754	0.770	0.754	0.770	0.770	0.705	0.770	0.787	0.717	0.726	0.590	0.623	0.721	0.721	0.393	0.574	0.371	0.608
		C2	0.113	0.704	0.310	0.423	0.521	0.507	0.634	0.606	0.704	0.746	0.634	0.690	0.901	0.846	0.904	0.423	0.535	0.789	0.676	0.549	0.549	0.556	0.373
		C3	0.113	0.704	0.310	0.423	0.521	0.507	0.634	0.606	0.704	0.746	0.634	0.690	0.901	0.846	0.904	0.423	0.535	0.789	0.676	0.549	0.549	0.407	0.582
	Bottom Right	C1	0.426	0.902	0.574	0.721	0.803	0.590	0.803	0.590	0.689	0.721	0.803	0.738	0.869	0.939	0.949	0.639	0.770	0.787	0.820	0.574	0.721	0.519	0.683
		C2	0.326	0.930	0.628	0.442	0.581	0.605	0.744	0.814	0.698	0.884	0.884	0.930	0.930	0.782	0.836	0.791	0.651	0.698	0.674	0.535	0.581	0.546	0.514
		C3	0.326	0.930	0.628	0.442	0.581	0.605	0.744	0.814	0.698	0.884	0.884	0.930	0.930	0.782	0.836	0.791	0.651	0.698	0.674	0.535	0.581	0.420	0.643

Table 12: Effect of the in-page corruption on the Page-Level Accuracy by varying in-context learning strategy and complexity level (addressing RQ2 and RQ3). DUDE dataset.

			Phi4		Molmo	Ovis		Llama		Llava 34B		Gemma 27B		Qwen 2.5 7B		Qwen 2.5 72B		InternVL 3 9B		InternVL 3 78B		GPT-4.1-mini		O3	
			Explicit	OCR Explicit	Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit	Explicit	OCR Explicit
Document Element	Title	C1	0.000	0.250	0.250	0.250	0.250	0.250	0.500	0.750	0.250	0.750	0.250	0.250	0.750	0.667	0.667	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
		C2	0.000	0.267	0.333	0.333	0.267	0.133	0.533	0.533	0.667	0.267	0.667	0.267	0.600	0.467	0.800	0.067	0.267	0.267	0.533	0.467	0.267	0.333	0.267
		C3	0.118	0.588	0.588	0.471	0.529	0.647	0.647	0.765	0.765	0.824	0.882	0.824	0.882	0.826	0.826	0.588	0.588	0.471	0.706	0.294	0.588	0.294	0.412
	Text	C1	0.212	0.715	0.676	0.782	0.559	0.721	0.788	0.726	0.715	0.844	0.782	0.844	0.894	0.873	0.891	0.709	0.788	0.765	0.827	0.704	0.760	0.665	0.480
		C2	0.193	0.665	0.628	0.642	0.693	0.647	0.697	0.775	0.688	0.789	0.711	0.835	0.853	0.778	0.856	0.679	0.720	0.656	0.803	0.583	0.725	0.624	0.550
		C3	0.232	0.580	0.536	0.723	0.732	0.527	0.661	0.607	0.625	0.750	0.696	0.768	0.804	0.828	0.873	0.518	0.545	0.786	0.839	0.527	0.723	0.491	0.348
	Figure	C1	0.258	0.774	0.613	0.387	0.355	0.677	0.677	0.710	0.710	0.774	0.935	0.839	0.806	0.667	0.688	0.645	0.645	0.645	0.806	0.581	0.677	0.581	0.484
		C2	0.070	0.721	0.581	0.326	0.535	0.674	0.651	0.721	0.698	0.512	0.674	0.791	0.884	0.623	0.721	0.651	0.744	0.628	0.791	0.372	0.628	0.279	0.279
		C3	0.364	0.364	0.455	0.364	0.364	0.455	0.545	0.545	0.364	0.636	0.545	0.727	0.818	0.818	0.818	0.273	0.364	0.455	0.636	0.364	0.273	0.273	0.273
	Table	C1	0.103	0.931	0.862	0.655	0.759	0.828	0.897	0.759	0.897	0.828	0.966	1.000	1.000	0.942	0.942	0.655	0.793	0.724	0.966	0.793	0.793	0.759	0.621
		C2	0.046	0.369	0.492	0.446	0.477	0.385	0.369	0.446	0.323	0.554	0.554	0.569	0.569	0.518	0.542	0.077	0.262	0.108	0.354	0.323	0.308	0.292	0.154
		C3	0.043	0.783	0.783	0.609	0.652	0.826	0.739	0.870	0.696	0.826	0.913	0.870	0.957	0.861	0.972	0.739	0.826	0.435	0.739	0.565	0.739	0.565	0.609
	Abandon	C1	0.471	0.941	0.824	0.882	0.000	0.882	0.824	0.647	0.824	0.882	0.882	0.882	0.941	1.000	1.000	0.706	0.824	0.941	0.882	0.882	0.765	0.706	0.176
		C2	0.300	0.700	0.600	0.400	0.500	0.600	0.800	1.000	0.800	0.700	0.500	0.700	0.700	0.600	1.000	0.600	0.600	0.700	0.700	0.500	0.500	0.400	0.300
		C3	0.667	0.667	1.000	0.667	0.667	1.000	1.000	0.667	1.000	1.000	0.667	0.667	1.000	0.667	0.667	0.667	1.000	1.000	1.000	0.667	0.667	0.667	0.667
Layout	Top Left	C1	0.152	0.780	0.644	0.674	0.682	0.689	0.765	0.780	0.788	0.833	0.818	0.818	0.856	0.857	0.883	0.644	0.765	0.614	0.780	0.712	0.735	0.652	0.500
		C2	0.102	0.408	0.453	0.445	0.524	0.427	0.455	0.654	0.448	0.584	0.647	0.599	0.610	0.526	0.613	0.298	0.416	0.259	0.497	0.411	0.469	0.387	0.296
		C3	0.102	0.408	0.453	0.445	0.524	0.427	0.455	0.654	0.448	0.584	0.647	0.599	0.610	0.526	0.613	0.298	0.416	0.259	0.497	0.411	0.469	0.387	0.296
	Top Right	C1	0.420	0.820	0.760	0.760	0.520	0.780	0.900	0.760	0.860	0.900	0.880	0.920	0.940	0.911	0.924	0.760	0.800	0.840	0.900	0.800	0.840	0.820	0.740
		C2	0.324	0.794	0.735	0.696	0.716	0.765	0.784	0.637	0.716	0.784	0.706	0.882	0.902	0.827	0.860	0.608	0.745	0.588	0.794	0.627	0.667	0.569	0.578
		C3	0.324	0.794	0.735	0.696	0.716	0.765	0.784	0.637	0.716	0.784	0.706	0.882	0.902	0.827	0.860	0.608	0.745	0.588	0.794	0.627	0.667	0.569	0.578
	Bottom Left	C1	0.347	0.810	0.777	0.793	0.421	0.777	0.785	0.686	0.752	0.826	0.818	0.884	0.934	0.918	0.927	0.736	0.818	0.793	0.843	0.760	0.818	0.678	0.364
		C2	0.096	0.664	0.528	0.504	0.600	0.624	0.608	0.728	0.616	0.600	0.592	0.728	0.800	0.685	0.751	0.600	0.616	0.512	0.704	0.424	0.584	0.576	0.488
		C3	0.096	0.664	0.528	0.504	0.600	0.624	0.608	0.728	0.616	0.600	0.592	0.728	0.800	0.685	0.751	0.600	0.616	0.512	0.704	0.424	0.584	0.576	0.488
	Bottom Right	C1	0.119	0.712	0.814	0.678	0.441	0.847	0.797	0.695	0.627	0.864	0.847	0.949	0.949	0.832	0.832	0.780	0.797	0.881	0.915	0.729	0.729	0.729	0.475
		C2	0.229	0.780	0.789	0.651	0.716	0.789	0.826	0.752	0.734	0.771	0.743	0.908	0.936	0.878	0.900	0.716	0.844	0.789	0.917	0.651	0.807	0.615	0.615
		C3	0.229	0.780	0.789	0.651	0.716	0.789	0.826	0.752	0.734	0.771	0.743	0.908	0.936	0.878	0.900	0.716	0.844	0.789	0.917	0.651	0.807	0.615	0.615

Table 13: Effect of the in-page corruption on the Page-Level Accuracy by varying in-context learning strategy and complexity level (addressing RQ2 and RQ3). MPDocVQA dataset.