AgenticOCR: Parsing Only What You Need for Efficient Retrieval-Augmented Generation

Zhengren Wang[1,2*], **Dongsheng Ma**[2,3*], **Huaping Zhong**[4*], **Jiayu Li**[2]
Yijie Wang[2], **Tianyao He**[2], **Wentao Zhang**[2,3], **Bin Wang**[2,1†], **Conghui He**[2,1†]
[1]School of Artificial Intelligence, Shanghai Jiao Tong University
[2]Shanghai Artificial Intelligence Laboratory [3]Peking University [4]SenseTime
ictwangbin@gmail.com heconghui@pjlab.org.cn

Abstract

The expansion of retrieval-augmented generation (RAG) into multimodal domains has intensified the challenge for processing complex visual documents, such as financial reports. While page-level chunking and retrieval is a natural starting point, it creates a critical bottleneck: delivering entire pages to the generator introduces excessive extraneous context. This not only overloads the generator's attention mechanism but also dilutes the most salient evidence. Moreover, compressing these information-rich pages into a limited visual token budget further increases the risk of hallucinations. To address this, we introduce AgenticOCR, a dynamic parsing approach that extends optical character recognition (OCR) from a static, full-text process into a query-driven, on-demand extraction system. By autonomously analyzing document layout in a "thinking with images" manner, AgenticOCR identifies and selectively recognizes regions of interest. This approach performs on-demand decompression of visual tokens precisely where needed, effectively decoupling retrieval granularity from rigid page-level chunking. AgenticOCR serves as a trainable evidence parser between page retrieval and answer generation in visual document RAG. Experimental results demonstrate that AgenticOCR improves both the accuracy and efficiency of visual RAG systems with pixel-adaptive per-image token allocation (e.g., Qwen3.5 generator), achieving expert-level performance in long document understanding. Our code is available at https://github.com/opendatalab/AgenticOCR.

1 Introduction

Optical character recognition (OCR) has evolved into a remarkably mature field. Driven by advances in vision-language models (VLMs), billion-parameter models such as MinerU2.5-Pro (1.2B) and PaddleOCR-VL-1.5 (0.9B) routinely achieve ∼95% accuracy on OmniDocBench [26, 36], while even ultra-lightweight variants like OpenOCR (0.1B) reach the 90% threshold. These advances provide a strong foundation for full-document parsing, while query-dependent evidence extraction remains less developed.

Yet this very success has created a paradox: while the supply of general-purpose, full-document parsing capability nears saturation, a critical demand gap for instruction-following or on-demand parsing remains unmet in downstream applications, especially in retrieval-augmented generation (RAG) over visual documents such as financial reports, technical manuals, and scholarly articles.

In visual RAG pipelines, documents are typically chunked and retrieved at the page level—a natural but coarse granularity [38, 44]. It delivers entire document pages—often containing headers, footers,

*Equal contribution. †Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

decorative elements, and irrelevant sections—to the generator. This mismatch between information need and retrieval granularity introduces two compounding inefficiencies. First, the generator's attention mechanism becomes diluted by extraneous visual context, reducing its capacity to focus on query-relevant evidence. Second, and more critically, high-resolution document pages must be compressed into limited visual token budgets, inevitably sacrificing fine-grained details and amplifying hallucination risks—particularly for rotated tables, small-font annotations, intricate formulas or complex layouts. Furthermore, recent studies on multi-image reasoning highlight that VLMs struggle with evidence localization in visual RAG and are easily misled by irrelevant visual information [31, 38], underscoring the inherent limitations of page-level retrieval.

This challenge cannot be resolved by further improving full-page OCR accuracy alone, but motivates extending "parsing everything" (base model) toward "parsing only what you need" (instruction-following model). Here, we introduce AgenticOCR: a dynamic, query-driven approach that extends OCR from a static pre-processing step toward an agentic process. By autonomously reasoning over document layout and content, an AgenticOCR model identifies, locates, and selectively parses the regions of interest relevant to a specific query or instruction. As shown in Fig. 1, with zoom, rotate and other operations, it performs on-demand decompression of visual information precisely where needed. Meanwhile, acting as a lightweight filter, it can serve as a computationally efficient pre-processor for large generative models like Gemini and GPT, reducing reliance on frontier models for evidence extraction. With pixel-adaptive per-image token allocation (Qwen3.5), Evidence+OCR reduces generator input tokens compared with Page+OCR. To realize this approach, we develop a scalable data synthesis pipeline that automatically generates training trajectories with element-level bounding box annotations, and train lightweight 4B/8B models through a two-stage framework combining trajectory distillation and reinforcement learning with spatially-aware rewards.

Contributions Our main contributions are threefold:

- **Structured Evidence Parsing**: We develop a typed OCR interface that returns element-level bounding boxes and recognized evidence for downstream generation, unifying layout analysis, geometric correction, and direct recognition.
- **Model Realization**: We develop an automated data synthesis pipeline that generates diverse training samples with consistent element-level bounding box annotations, yielding 9.5K SFT trajectories and 20K+ RL samples. With a two-stage training framework (rejection sampling from Gemini-3.1-Pro-Preview followed by GRPO with specialized rewards), our pipeline produces lightweight but competitive 4B/8B models as open-sourced research prototypes.
- **Empirical Validation**: Our experiments demonstrate that integrating AgenticOCR into visual RAG pipelines surpasses the human expert performance on MMLongBench-Doc and rivals heavyweight multi-agent systems such as DocLens [53] and SimpleDoc [14]. We also discuss current limitations and future optimization directions to inspire further research.

2 Related work

Visual RAG and document parsing Retrieval-Augmented Generation (RAG), initially successful in textual domains, has been extended to handle visually rich documents [1, 17]. Early visual RAG systems such as ColPali [8] and VisRAG [44] pioneered the use of document page snapshots as retrieval units, directly feeding images to VLMs to preserve layout semantics. However, this page-level granularity introduces substantial irrelevant visual content, diluting the generator's attention and increasing hallucination risk [32, 37]. Recent efforts refine this granularity: RegionRAG [20] shifts retrieval to the region level using contrastive learning and post-processing to group relevant patches; DocLens [53] employs multi-agent systems for hierarchical evidence localization; DocDancer [48] formulates DocVQA as an agentic information-seeking problem with search and read tools; VRAG-RL [38] and EVisRAG [31] integrate reinforcement learning to interleave retrieval, visual perception, and reasoning. On the parsing side, traditional pipelines [47, 35] decompose parsing into sequential modules, offering interpretability at the cost of error propagation. End-to-end VLM architectures [3, 25, 40, 28] eliminate cascaded failures but introduce native-resolution $\mathcal{O}(N^2)$ complexity and token redundancy from blank regions. Recent multi-stage methods [9, 21, 6] address this by decoupling layout analysis from content recognition—first analyzing downsampled pages globally, then parsing cropped regions locally—exemplified by MinerU2.5's coarse-to-fine mechanism [24, 36].

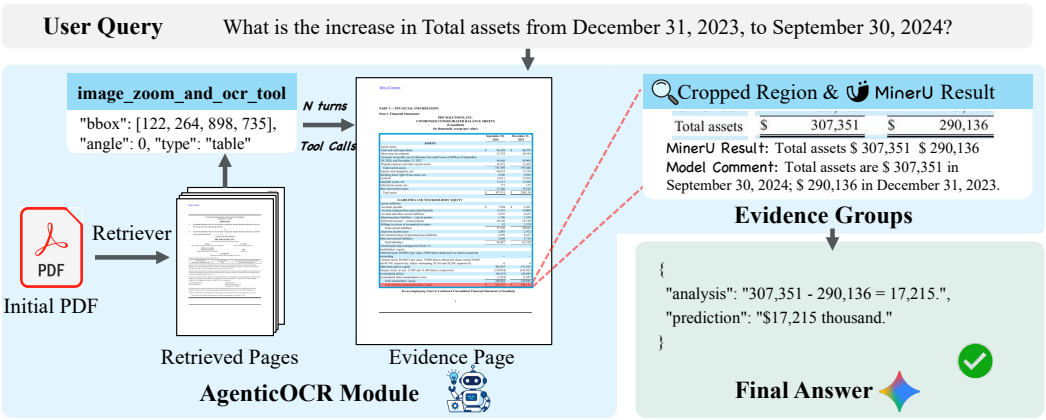


Figure 1: Overview of AgenticOCR-based visual RAG. Given a query and retrieved document pages, AgenticOCR autonomously analyzes relevance, identifies query-relevant regions, and invokes the `image_zoom_and_ocr_tool` with operations such as zoom and rotate to perform on-demand decompression of visual information. The extracted evidence (cropped image and OCR text) is then forwarded to the answer generator, filtering irrelevant context before answer generation.

Furthermore, PaddleOCR-VL-1.5 [7] and HunyuanOCR [34] enhance robustness against physical distortions. Meanwhile, DeepSeek-OCR [41] treats visual tokens as a compression medium, achieving 7–20$\times$ token reduction while preserving decoding fidelity; its successor [42] further models visual causal flow through LLM-based encoders. Despite achieving $\sim$95% accuracy on OmniDocBench, all these systems remain static—processing everything regardless of downstream need—and specialized models for on-demand, query-driven evidence extraction remain largely underexplored.

Agentic multimodal models Recent work on multimodal reasoning has explored a shift from passive perception toward more active interaction, where models dynamically engage with visual inputs through tool invocation. OpenAI's o3 [30] introduced the concept of "thinking with images", demonstrating that interleaving visual manipulation with textual reasoning enables more faithful visual cognition. DeepEyes [52] incentivizes region-focused visual exploration via an image zoom-in tool, allowing models to autonomously crop regions of interest during reasoning without cold-start supervision. Thyme [49], PyVision [50] and CodeVision [11] expand functionality by enabling code-based image operations (e.g., rotation, contrast adjustment, multi-step transformations) through SFT+RL training. DeepEyesV2 [13] explores integrating external knowledge acquisition via online image search, recognizing that perception alone is insufficient for knowledge-intensive tasks. Despite these advances, existing agentic frameworks remain oriented toward general visual reasoning without addressing the parsing demands of document intelligence. AgenticOCR bridges this gap.

Scope and distinction. AgenticOCR specializes a small VLM into a structured evidence parser through a typed OCR interface, element-level supervision, and spatial and behavioral rewards. Region-RAG learns contrastive region retrieval; DocLens combines frontier-model navigation, localization, and answer validation; general tool-using VLMs primarily produce answers. AgenticOCR returns OCR text and bounding boxes that can be reused by downstream generators. Its page-wise parsing complements global navigation and multi-hop reasoning. Appendix A provides further comparisons.

3 AgenticOCR: parsing only what you need

We present AgenticOCR, a dynamic parsing approach that extends OCR from a static pre-processing step toward a query-driven, agentic process. Our approach includes three core components: (1) a unified visual interaction tool that enables fine-grained, on-demand evidence extraction; (2) a scalable data synthesis pipeline and SFT+RL training framework that produces lightweight but competitive models; (3) an integration protocol that allows AgenticOCR to serve as a plug-and-play module within visual RAG pipelines. Fig. 1 illustrates the overall workflow.

3.1 The `image_zoom_and_ocr_tool` primitive

For flexible parsing, we design a single but versatile tool that unifies region localization, geometric correction, and content recognition into an atomic operation. Formally, the tool function is defined as:

$$\mathcal{T} : (I, \mathbf{b}, \theta, \tau) \mapsto (I_{\text{crop}}, \mathcal{R}), \tag{1}$$

where I is the input page image; $\mathbf{b} = [x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ denotes a bounding box in normalized coordinates (0–1000) relative to the original page image, specifying the region to crop; $\theta \in \{0°, 90°, 180°, 270°\}$ specifies rotation angle applied counter-clockwise to the cropped region; and $\tau \in \{\texttt{region}, \texttt{text}, \texttt{table}, \texttt{image}, \texttt{equation}\}$ indicates the semantic type of the cropped images. The tool returns both the cropped image I_{crop} and structured recognition results $\mathcal{R}$. Crucially, the tool's behavior depends on the specified type τ:

- **Region mode ($\tau = region$):** Invokes MinerU's full two-stage pipeline—first performing layout analysis on the zoomed-in region, then executing fine-grained recognition on detected sub-elements. This mode is designed to **enhance active perception for complex regions**, splitting regions of interest from the whole page. Notably, the tool returns layout results with coordinates normalized to the original page rather than the cropped image. This circumvents confusion between different coordinate systems and facilitates the tracing of evidence.
- **Element mode ($\tau \in \{text, table, equation\}$):** Skips layout detection and directly applies MinerU's recognition capability to the cropped region. This bypasses layout detection to directly parse target elements, improving efficiency and reducing cascading errors.
- **Image mode ($\tau = image$):** Returns only the cropped visual patch without OCR, enabling pure visual perception before committing to content extraction.

3.2 Data synthesis and model training

Scalable data synthesis pipeline. Training AgenticOCR models requires large-scale trajectories with element-level bounding box annotations—data that is expensive to annotate manually and prone to inconsistent granularity across human annotators. To address this, we develop a fully automated synthesis pipeline. Starting from a large pool of PDF documents sourced from Common Crawl[1], we first apply VLM-based domain annotation and stratified sampling to ensure broad coverage. Each selected PDF is then parsed by MinerU2.5 [24, 36] to extract fine-grained structural elements with bounding box coordinates. High-performance VLMs then serve as intelligent agents that navigate the parsed bounding box space across pages, identifying and aggregating semantically related elements into comprehensive Evidence Packages—each containing multi-page, multi-element evidence chains with precise spatial coordinates. Based on these Evidence Packages, we synthesize QA pairs through template-driven generation. Finally, a multi-stage quality control process—including answerability verification and relevance filtering—ensures data quality. The full pipeline details are provided in Appendix C. This pipeline produces training data with more consistent annotation granularity than human-labeled alternatives, and scales to generate both SFT trajectories and large-scale RL samples.

Cold-start via supervised fine-tuning. Direct RL training often fails due to sparse rewards and unstable exploration [13]. We therefore employ a cold-start phase using high-quality trajectories distilled from Gemini via rejection sampling [39, 48]. Our goal is not maximal performance but to instill a stable prior for when and how to invoke the tool.

We construct SFT trajectories from two sources: the ViDoRe-v3 benchmark [22], which contains queries, document pages, and human-annotated bounding boxes $\mathcal{B}^{\text{gt}}$ for relevant regions, and our automated synthesis pipeline, which generates diverse samples with consistent element-level annotations. We employ rejection sampling on Gemini-3.1-Pro-Preview [33] to obtain high-quality, multi-turn reasoning traces. Since bbox's annotations may exhibit inconsistent granularity, we apply a dual-threshold filtering strategy. The coverage and exact-match recalls used for filtering are defined below with the reward. We first filter trajectories requiring $\text{Recall}_{\min} \geq 0.8$ to ensure coverage, then rank by $\text{Score} = \text{Recall}_{\min} + \text{Recall}_{\text{EM}}$ to prioritize spatial precision. This yields 8K high-quality positive trajectories (approximately 30% from ViDoRe-v3 and 70% from our synthesis pipeline).

[1] `https://commoncrawl.org/`

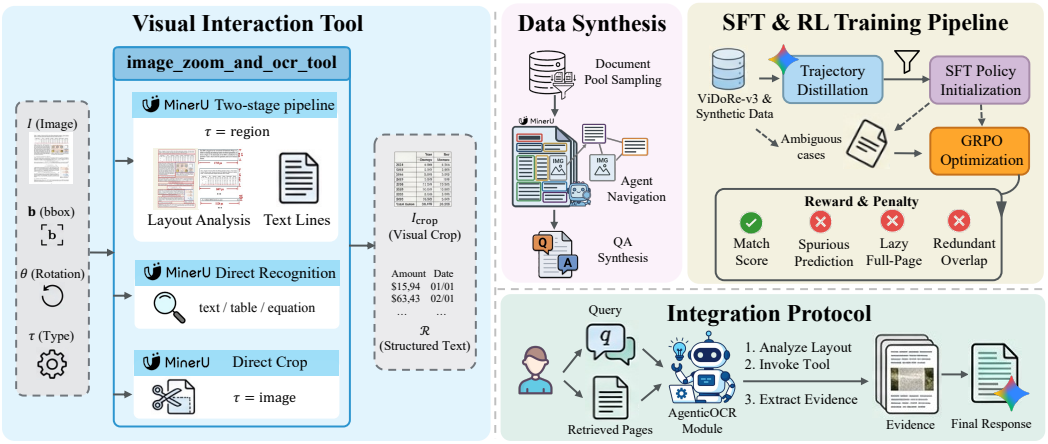


Figure 2: AgenticOCR architecture and training pipeline. **Left**: The Visual Interaction Tool employs MinerU for document parsing, applying visual cropping or OCR based on the element type. **Top Right**: The SFT policy is initialized via trajectory distillation on ViDoRe and our synthetic data. Ambiguous cases undergo GRPO with a reward system that favors accurate matches while penalizing spurious, redundant, or lazy full-page predictions. **Bottom Right**: At inference, AgenticOCR extracts evidence from retrieved pages for user query, which is then forwarded to the answer generator.

Notably, the synthetic data provides more consistent bounding box granularity than human annotations, reducing noise in the training signal. To suppress spurious tool calls on irrelevant pages, we additionally construct 1.5K hard negative samples from ViDoRe-v3 by selecting semantically similar but logically unrelated pages via rerank score and Gemini verification (details in Appendix D.2), giving 9.5K SFT samples in total.

Alignment via reinforcement learning. While SFT establishes basic tool-use competence, it cannot optimize for spatial precision or human preference alignment. We apply Group Relative Policy Optimization (GRPO) [29] with a tailored reward function that jointly optimizes coverage, localization accuracy, and behavioral constraints.

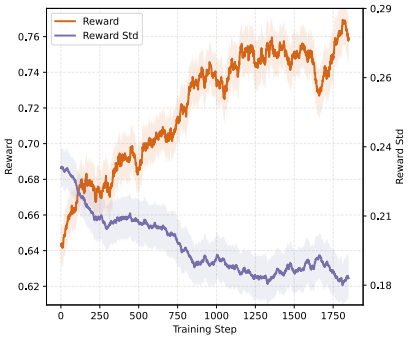


Figure 3: GRPO reward mean and std.

To maximize training effectiveness, we leverage our synthesis pipeline to generate a large-scale RL dataset of over 20k samples spanning diverse domains, layouts, and question types—including rotated and complex layouts. We apply uncertainty-based difficulty filtering: for each candidate, we collect 8 rollouts from the SFT model and retain samples whose standard deviation of the accuracy reward R_{acc} (Eq. 5) exceeds 0.15. This concentrates optimization on the most informative examples, avoiding wasted computation on already-converged easy cases or hopelessly difficult outliers. Negative samples are included at $\sim$20% ratio to reinforce page-level relevance judgment.

Reward design. For positive samples, we reward two complementary aspects of evidence localization: covering the relevant content and matching its spatial extent. The first tolerates differences in annotation granularity; the second favors tightly aligned boxes. Both are computed on the final JSON evidence boxes returned by the policy.

Given ground-truth boxes $\mathcal{B}^{\text{gt}}$ and predicted boxes $\mathcal{B}^{\text{pred}}$, we define:

$$\text{IoU}_{\text{EM}}(b^{\text{gt}}, b^{\text{pred}}) = \frac{|b^{\text{gt}} \cap b^{\text{pred}}|}{|b^{\text{gt}} \cup b^{\text{pred}}|}, \quad \text{IoU}_{\min}(b^{\text{gt}}, b^{\text{pred}}) = \frac{|b^{\text{gt}} \cap b^{\text{pred}}|}{\min(|b^{\text{gt}}|, |b^{\text{pred}}|)}, \tag{2}$$

$$\mathrm{Recall}_{\mathrm{EM}} = \frac{1}{|\mathcal{B}^{\mathrm{gt}}|} \sum_{b^{\mathrm{gt}} \in \mathcal{B}^{\mathrm{gt}}} \mathbb{I}\Big(\max_{b^{\mathrm{pred}} \in \mathcal{B}^{\mathrm{pred}}} \mathrm{IoU}_{\mathrm{EM}}(b^{\mathrm{gt}}, b^{\mathrm{pred}}) \geq thres_{\mathrm{EM}} \Big), \tag{3}$$

$$\mathrm{Recall}_{\mathrm{min}} = \frac{1}{|\mathcal{B}^{\mathrm{gt}}|} \sum_{b^{\mathrm{gt}} \in \mathcal{B}^{\mathrm{gt}}} \mathbb{I}\Big(\max_{b^{\mathrm{pred}} \in \mathcal{B}^{\mathrm{pred}}} \mathrm{IoU}_{\mathrm{min}}(b^{\mathrm{gt}}, b^{\mathrm{pred}}) \geq thres_{\mathrm{min}} \Big), \tag{4}$$

where $\mathrm{IoU}_{\mathrm{min}}$ uses the minimum box area as denominator (tolerant to granularity mismatch) and $\mathrm{IoU}_{\mathrm{EM}}$ is the standard intersection-over-union (sensitive to precise alignment). We adopt $thres_{\mathrm{EM}} = 0.6$ and $thres_{\mathrm{min}} = 0.8$.

The accuracy reward averages these two recalls:

$$R_{\mathrm{acc}} = \tfrac{1}{2}(\mathrm{Recall}_{\mathrm{min}} + \mathrm{Recall}_{\mathrm{EM}}). \tag{5}$$

A prediction contained within a larger annotated region can receive coverage credit through $\mathrm{Recall}_{\mathrm{min}}$, while $\mathrm{Recall}_{\mathrm{EM}}$ encourages agreement with the full region. Combining them supplies a useful learning signal across coarse and fine annotations and favors accurate localization.

Inspired by ZwZ [43], we also reward accurate localization without unnecessary tool use. For a trajectory with T tool calls, the adaptive bonus is

$$B_{\mathrm{adapt}} = 0.1\,\mathbb{I}\big[T = 0 \ \wedge \ R_{\mathrm{acc}} \geq 0.8\big]. \tag{6}$$

The model receives this bonus when it localizes the evidence accurately from the page image alone. Tool-free trajectories still return evidence boxes and use the same accuracy calculation.

For positive samples, the complete reward is

$$R_{\mathrm{pos}} = R_{\mathrm{acc}} + B_{\mathrm{adapt}} - P_{\text{over-pred}} - P_{\mathrm{overlap}} - P_{\mathrm{oversize}}. \tag{7}$$

The three penalties discourage spurious boxes, redundant overlaps, and oversized `region` tool calls; their definitions are given in Appendix D.3. For negative samples, which contain no query-relevant evidence, we use $R_{\mathrm{neg}} = 1$ if the final box set is empty and 0 otherwise, reinforcing page-level relevance judgment.

3.3 Integration with visual RAG pipelines

AgenticOCR operates as a plug-and-play middleware between page retrieval and generation modules (Figure 1). Given a query q and top-k retrieved pages from visual page retrievers (e.g., Qwen3-VL-Reranker [19]), AgenticOCR independently processes each page: analyzing layout and query semantics, invoking `image_zoom_and_ocr_tool` to extract structured evidence $\mathcal{E}_i = \{(b_j, e_j)\}_j$ with boxes b_j and evidence text e_j, and formatting results as interleaved text–image snippets for the generator. This preserves visual context and requires no modification to existing generators.

Table-aware OCR fallback. When the extracted evidence on a page contains a table element, we retain the localized visual crops but replace that page's localized OCR and model comments with full-page OCR text. Non-table pages remain unchanged. This preserves headers, units, and surrounding table context that a fragmented crop or generated comment can omit.

4 Experiments

4.1 Experimental setup

Benchmarks and metrics. We evaluate our visual RAG system on two challenging benchmarks: MMLongBench-Doc [23] and FinRAGBench-V [51]. MMLongBench-Doc tests reasoning over lengthy, multi-domain documents (avg. 49.4 pages) that require integrating scattered information across diverse modalities. The dedicated "Unanswerable" subset evaluates hallucination response. Performance of human experts on this benchmark is reported as 65.8. FinRAGBench-V is helpful for our analysis due to two unique features: its use of documents with dense, newspaper-like layouts, and its support for evaluating visual citation (block- or element-level). We adhere to original evaluation protocols: rule-based scoring for MMLongBench-Doc and an LLM-as-a-judge for FinRAGBench-V.

Table 1: Main results on MMLongBench-Doc and FinRAGBench-V. We report the accuracy of five types of evidence sources including pure text (TXT), layout (LAY), chart (CHA), table (TAB), and figure (FIG), and on unanswerable (UNA) samples. **Bold** and underlined indicate the best and the second score per column. † denotes results reported in original papers, hence some results are unavailable. * Denotes results surpassing human experts (65.8 on MMLongBench-Doc).

Method	MMLongBench-Doc							FinRAGBench-V			
	TXT	LAY	CHA	TAB	FIG	UNA	ALL	TXT	TAB	CHA	ALL
Vanilla VLMs											
GPT-4o†	46.3	46.0	45.3	50.0	44.1	20.2	42.8	-	-	-	37.2
Claude-4-Sonnet	50.4	49.4	50.5	57.3	43.9	59.0	53.4	36.6	20.2	51.9	33.8
Gemini-2.5-Flash	44.0	53.2	46.0	43.9	48.2	56.7	49.6	49.0	41.6	41.0	43.0
Gemini-2.5-Pro	52.1	62.1	55.5	55.3	54.0	59.9	58.1	62.2	55.3	50.4	54.9
o4-mini†	-	-	-	-	-	-	-	-	-	-	62.4
VLMs Augmented with OCR											
Claude-4-Sonnet	52.7	51.6	50.0	58.1	45.3	65.9	56.0	58.7	21.6	54.3	41.0
Gemini-2.5-Flash	55.9	54.9	52.7	63.4	50.3	60.8	58.5	67.6	64.4	46.1	58.3
Gemini-2.5-Pro	59.7	65.3	60.8	68.3	55.7	58.4	63.3	70.0	70.0	56.2	64.9
VLM-based Agentic Frameworks											
MACT (w/ MiMo-VL-7B)†	-	-	-	-	-	-	47.4	-	-	-	-
M3DocRAG (w/ Qwen2-VL-7B)†	30.0	23.5	18.9	20.1	20.8	5.8	21.0	-	-	-	-
MDocAgent (w/ GPT-4o)†	-	-	-	-	-	-	42.0	-	-	-	-
SimpleDoc											
w/ Claude-4-Sonnet	52.1	53.3	58.3	62.4	46.9	66.5	58.6	59.6	68.9	54.9	61.7
w/ Gemini-2.5-Flash	45.5	57.4	49.0	51.6	45.2	66.5	53.3	70.2	56.2	53.6	58.3
w/ Gemini-2.5-Pro	48.4	54.8	55.7	56.1	52.5	59.7	56.6	67.5	64.0	60.9	63.6
DocLens											
w/ Claude-4-Sonnet	59.9	58.2	54.4	63.9	55.3	**74.0**	63.3	70.2	66.0	60.3	64.8
w/ Gemini-2.5-Flash	59.5	61.5	54.8	66.9	59.0	73.8	64.7	69.9	71.3	64.5	68.5
w/ Gemini-2.5-Pro	63.7	64.6	**64.3**	**69.7**	60.2	72.2	**67.6***	68.9	74.2	67.1	70.4
AgenticOCR (Ours)											
AgenticOCR-4B *(Gemini-2.5-Pro generator)*											
w/ Page	57.9	64.5	59.5	57.3	58.0	63.3	61.6	85.4	68.5	65.7	71.1
w/ Page + OCR	62.8	67.3	58.9	64.4	57.8	66.0	64.1	89.1	**77.4**	66.8	76.0
w/ Evidence	61.6	66.6	63.8	62.8	63.1	63.2	64.7	86.8	75.5	73.5	77.2
w/ Evidence + OCR	62.4	66.3	62.5	62.3	62.6	66.6	64.9	89.0	74.8	73.4	77.4
AgenticOCR-8B *(Gemini-2.5-Pro generator)*											
w/ Page	57.5	60.7	56.1	58.4	55.4	66.3	60.9	85.1	69.9	66.0	71.8
w/ Page + OCR	63.3	62.2	57.4	67.0	58.2	70.6	65.0	**89.4**	76.9	68.5	76.5
w/ Evidence	64.2	67.2	61.5	62.9	61.2	66.7	64.9	89.1	76.4	73.7	78.2
w/ Evidence + OCR	**67.4**	**68.8**	63.1	64.3	**63.6**	66.8	66.4*	88.7	76.2	**75.3**	**78.6**

Models and baselines. We train two variants of our proposed modules: AgenticOCR-4B and AgenticOCR-8B, initialized from Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct respectively. Following DocLens, the performance is benchmarked against three categories of baselines. (1) **Vanilla VLMs**: Models that receive only page screenshots without any additional processing. For comparison and ablations, this includes Claude-4-Sonnet, Gemini-2.5-Flash, Gemini-2.5-Pro and so on; (2) **VLMs augmented with OCR**: The same VLMs are provided with both page screenshots and OCR-extracted text appended to each page; (3) **VLM-based Agentic Frameworks**: We compare with MACT [45], M3DocRAG [5], MDocAgent [12], SimpleDoc [14], and DocLens [53]. Performance numbers for MACT, M3DocRAG, and MDocAgent are taken from best results in original papers.

Further statistics, training and evaluation details are provided in Appendices B and E.

4.2 Main results

Table 1 presents the main results on both benchmarks. Our proposed AgenticOCR achieves competitive performance across the board, with several key observations.

Overall performance. On MMLongBench-Doc, our 8B model with *Evidence+OCR* achieves 66.4 overall accuracy, surpassing human experts (65.8) and approaching DocLens with Gemini-2.5-Pro (67.6). On FinRAGBench-V, it reaches 78.6, outperforming prior agentic frameworks. Notably,

Table 2: Table-subset accuracy with AgenticOCR-8B and a Gemini-2.5-Pro generator under table-aware OCR fallback. The fallback retains local visual crops and substitutes full-page OCR for local OCR/comments only on pages containing extracted table elements.

	Page	Evidence	Page+OCR	Evidence+OCR	Fallback	DocLens
MMLongBench TAB	58.4	62.9	67.0	64.3	**70.2**	69.7
FinRAGBench TAB	69.9	76.4	76.9	76.2	**77.8**	74.2

AgenticOCR achieves this using only a 8B model for evidence extraction, whereas DocLens must first feed the entire PDF document with full OCR text to Gemini for 8 rounds of page navigation (taking the union of selected pages), then re-feed selected pages for element-level localization, and finally sample 8 candidate answers with an adjudicator for cross-validation—requiring repeated frontier-model calls across stages, with the total varying with the selected pages. SimpleDoc requires multi-stage iterative retrieval, summarization, and a separate consistency evaluation step, still powered by frontier models. See Appendix A for more comparison.

Strengths on text, layout, and figure modalities. AgenticOCR exhibits particularly strong performance on text (TXT), layout (LAY), and figure (FIG) subsets of MMLongBench-Doc. Our 8B model with *Evidence+OCR* reaches 67.4 on TXT, 68.8 on LAY, and 63.6 on FIG—surpassing the best DocLens configurations on all three. This suggests that the agent's ability to precisely localize relevant text blocks, recognize layout structure, and interpret figures via on-demand zoom is especially effective for these modalities. The *Evidence+OCR* input format, which interleaves low-resolution page snapshots, cropped visual patches, and extracted text, provides a rich yet compact representation that benefits both layout-sensitive and content-heavy reasoning.

Limitations on tables and unanswerable questions. The original Evidence+OCR setting reaches 64.3 on MMLongBench-Doc TAB, below Page+OCR (67.0) and DocLens (69.7), although it outperforms DocLens on FinRAGBench-V TAB (76.2 vs. 74.2). Two factors help explain this difference. First, 51.4% of MMLongBench table questions require cross-page evidence, whereas 87.8% of FinRAGBench table questions use a single page. Second, mixed annotation granularity in ViDoRe-v3 can favor fragmented table crops that omit headers or units; generated OCR comments may also misalign columns or numerical scales. The visual Evidence variant improves over Page on MMLongBench TAB (62.9 vs. 58.4), but adding local OCR/comments yields only 1.4 points, compared with 8.6 points from full-page OCR. This motivates preserving visual crops while restoring page-level text context. Table-aware fallback raises TAB accuracy to 70.2 on MMLongBench-Doc and 77.8 on FinRAGBench-V, and overall MMLongBench-Doc accuracy to 67.6 (Table 2). For unanswerable questions, accuracy remains 66.8; distractor pages can make rejection more difficult.

4.3 Ablation studies and discussions

4.3.1 End-to-end ablation on evidence extraction

Evidence extraction baselines. To isolate the contribution of query-driven extraction, we fix the retriever (Qwen3-VL-Reranker), OCR engine (MinerU), and generator (Gemini-2.5-Pro), varying only the evidence extractor. We compare against four baselines: **Reranker-only** forwards the top-k retrieved pages together with their full-page OCR text to the generator without any extraction, serving as a reference baseline. Note that this is distinct from both the *Gemini-2.5-Pro+OCR* vanilla setting in Table 1 (which feeds the *entire* PDF plus full-document OCR) and the *Page* setting under AgenticOCR in Table 1 (which further narrows the top-k to a query-relevant subset); **RegionRAG** [20] applies region-level retrieval within each retrieved page to select relevant regions; **Gemini-3.1-Pro-Preview** directly prompts the frontier-scale Gemini to identify evidence; **Gemini-3.1-Pro-Preview+Tool** equips the model with `image_zoom_and_ocr_tool`, serving as the most competitive baseline.

Several findings emerge from Table 3. First, the AgenticOCR variants improve overall accuracy over the Reranker-only baseline, confirming that filtering irrelevant visual context before generation is beneficial. Second, RegionRAG underperforms even the no-extraction baseline, indicating that ColPali-style region retrieval without reasoning can discard useful context. Third, GRPO yields substantial gains over SFT-only (66.4 vs. 63.8 on MMLongBench-Doc, 78.6 vs. 75.3 on

Table 3: End-to-end ablation on evidence extractor with fixed retriever (Qwen3-VL-Reranker), OCR engine (MinerU), generator (Gemini-2.5-Pro), varying only the evidence extractor.

Evidence Extractor	MMLongBench-Doc							FinRAGBench-V			
	TXT	LAY	CHA	TAB	FIG	UNA	ALL	TXT	TAB	CHA	ALL
Baseline Extractors											
Reranker-only	60.2	65.0	61.1	67.5	55.4	58.8	63.5	70.5	69.2	57.0	65.3
RegionRAG [20]	58.5	62.3	58.2	64.8	53.1	56.5	61.2	67.2	65.8	53.5	62.0
Gemini-3.1-Pro-Preview	63.5	65.8	61.5	65.2	59.8	63.2	64.5	86.2	72.5	71.2	75.2
Gemini-3.1-Pro-Preview+Tool	65.0	66.7	62.5	66.0	61.2	64.7	65.7	87.5	74.2	72.5	76.5
AgenticOCR (Ours)											
AgenticOCR-4B (SFT+GRPO)	62.4	66.3	62.5	62.3	62.6	66.6	64.9	89.0	74.8	73.4	77.4
AgenticOCR-8B (SFT+GRPO)	67.4	68.8	63.1	64.3	63.6	66.8	66.4	88.7	76.2	75.3	78.6
AgenticOCR-8B (SFT-only)	64.4	65.1	60.3	61.8	60.4	64.7	63.8	86.5	72.8	71.2	75.3
AgenticOCR-8B (SFT+GRPO, w/o all penalties)	66.1	67.2	62.3	63.5	62.8	65.4	65.3	88.2	75.1	73.8	77.2

Table 4: Intrinsic evidence extraction on FinRAGBench-V, MVToolbench (rotations reverted), and FinRAGBench-V hard negatives. P_{acc} is page accuracy; R/P denote recall/precision. Gemini denotes Gemini-3.1-Pro-Preview. Entries average 8 runs at temperature 1.0; see Table 10 for standard deviations.

Evidence Extractor	FinRAGBench-V (bbox)					MVToolbench					FinRAG Neg.
	$\mathbf{P_{acc}}$	$\mathbf{R_{min}}$	$\mathbf{P_{min}}$	$\mathbf{F1_{min}}$	$\mathbf{R_{EM}}$	$\mathbf{P_{acc}}$	$\mathbf{R_{min}}$	$\mathbf{P_{min}}$	$\mathbf{F1_{min}}$	$\mathbf{R_{EM}}$	$\mathbf{P_{acc}}$
Baseline Extractors											
RegionRAG [20]	100.0	98.0	55.6	63.5	7.7	60.8	9.2	1.7	2.5	0.2	0.5
Gemini	96.8	85.7	78.0	79.6	46.7	94.2	75.8	73.5	74.6	38.3	100.0
Gemini+Tool	94.0	83.5	78.6	78.0	46.4	92.8	75.7	73.1	73.9	31.1	100.0
AgenticOCR (Ours)											
8B (SFT+GRPO)	96.7	84.9	76.0	78.2	45.9	98.7	71.7	68.5	69.5	37.3	95.3
4B (SFT+GRPO)	94.6	83.8	73.8	76.5	46.4	99.3	70.7	66.4	67.9	40.5	81.3
8B (SFT-only)	93.7	79.9	74.9	75.8	40.0	98.0	71.2	68.8	69.6	31.6	98.8
8B (w/o penalties)	95.5	88.5	59.5	68.0	45.9	98.0	79.3	57.8	64.5	36.9	95.0

FinRAGBench-V), and removing all penalty terms degrades performance to 65.3/77.2, supporting the benefit of the combined penalties for precise extraction. Finally, AgenticOCR-8B exceeds Gemini-3.1-Pro-Preview+Tool in overall accuracy on both benchmarks under this setup.

4.3.2 Intrinsic evaluation of evidence extraction

To isolate AgenticOCR's extraction capability from the end-to-end pipeline, we evaluate on single-page scenarios that assess both page-level relevance judgment and element-level localization. Benchmarks include FinRAGBench-V (subset with bounding boxes), MVToolbench [11], and hard negatives from FinRAGBench-V. Metric definitions and baseline configurations are in Appendix E.5.

Results. Table 4 summarizes extraction quality. On FinRAGBench-V, AgenticOCR-8B reaches 96.7 page accuracy—matching Gemini-3.1-Pro-Preview (96.8) and exceeding its tool-enabled variant (94.0)—with $\mathrm{F1}_{\min}$ of 78.2 and $\mathrm{Recall}_{\mathrm{EM}}$ of 45.9, while AgenticOCR-4B attains 94.6/76.5/46.4 at roughly half the parameters. On MVToolbench, both 4B/8B variants achieve near-perfect page accuracy ($\geq$98.7), clearly surpassing Gemini-3.1-Pro-Preview (94.2) and indicating strong generalization to diverse layouts. As a baseline, RegionRAG—even at its highest recommended bbox threshold (0.3)—collapses on hard negatives (0.5 on FinRAGBench-V) and yields low-precision boxes ($\mathrm{Prec}_{\min}$ 55.6/1.7). Ablations confirm that GRPO consistently improves $\mathrm{Recall}_{\mathrm{EM}}$ over SFT-only (45.9 vs. 40.0 on FinRAGBench-V; 37.3 vs. 31.6 on MVToolbench) and, crucially, that penalty terms prevent the precision collapse observed in the w/o-penalty variant ($\mathrm{Prec}_{\min}$: 76.0 vs. 59.5 on FinRAGBench-V). A qualitative example is shown in Figure 5 in the Appendix.

Table 5: Accuracy and token consumption under four input configurations, with fixed retriever and extractor (AgenticOCR-8B), varying the generator's input format. Evidence+OCR improves accuracy and reduces input tokens relative to Page+OCR with the evaluated Qwen generator.

Dataset	Generator	Setting	Accuracy	Avg. Input Token	Avg. Output Token
MMLongBench-Doc Page-Level Metrics: (Recall: 93.5, Precision: 28.0)	Gemini-2.5-Pro	Page	60.9	3377.313	1403.871
		Page+OCR	65.0	7613.253	1343.515
		Evidence	64.9	7904.886	1399.083
		Evidence+OCR	66.4	9476.468	1409.607
	Qwen3.5-397B-A17B	Page	67.2	12091.894	3104.802
		Page+OCR	67.2	17161.588	2963.324
		Evidence	68.4	14559.887	3134.409
		Evidence+OCR	68.7	16364.014	3094.298
FinRAGBench-V Page-Level Metrics: (Recall: 95.3, Precision: 34.8)	Gemini-2.5-Pro	Page	71.8	3042.460	1930.657
		Page+OCR	76.5	9289.288	1721.593
		Evidence	78.2	6474.760	1849.603
		Evidence+OCR	78.6	9090.601	1832.599
	Qwen3.5-397B-A17B	Page	81.8	8360.708	2696.191
		Page+OCR	83.0	15464.034	2613.310
		Evidence	82.9	9508.211	2673.940
		Evidence+OCR	83.8	11536.733	2747.607

4.3.3 Less is more: higher accuracy with fewer input tokens

Table 5 compares four generator input formats. With Qwen3.5-397B-A17B, replacing *Page+OCR* with *Evidence+OCR* improves accuracy from 67.2 to 68.7 on MMLongBench-Doc and from 83.0 to 83.8 on FinRAGBench-V, while reducing input tokens by 4.6% and 25.4%, respectively. Appendix F.1 explains the effect of per-image token allocation.

4.3.4 Model specialization and efficiency

Under the same tool interface, prompt, and evaluator, Qwen3-VL-8B+Tool achieves 16.0 page accuracy on the FinRAGBench-V bbox subset, compared with 96.7 for AgenticOCR. Invalid output formats account for many baseline errors. Within AgenticOCR, GRPO raises page accuracy from 93.7 to 96.7 over SFT-only training and reduces average tool calls from 1.2 to 0.7. The trained parser also reduces frontier-model usage in the full pipeline. Appendix G provides the same-size baselines, extraction throughput, and end-to-end token and latency comparisons.

5 Conclusion

AgenticOCR is a trainable, lightweight evidence parser for visual document RAG. Its typed tool interface, element-level data synthesis, and SFT+GRPO training support query-driven extraction between page retrieval and answer generation. Experiments show strong downstream accuracy and reduced frontier-model dependence, with additional input-token savings under pixel-adaptive allocation. Table-aware OCR fallback improves structural context, while cross-page navigation and answer validation remain complementary capabilities.

6 Limitations

AgenticOCR processes pages independently, limiting cross-page evidence selection and multi-hop reasoning. Table-aware fallback recovers local headers and context, but cross-page tables require broader navigation. For charts, generated comments may misread legends, axes, or values and propagate errors to the answer generator. Distractor pages also reduce accuracy on unanswerable questions. Synthetic training trajectories may retain teacher and parser biases despite filtering and spatial-reward supervision. These issues motivate cross-page navigation, better chart interpretation, and improved training data.

Acknowledgments and Disclosure of Funding

This work is supported by Shanghai Artificial Intelligence Laboratory.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[2] Tyler Bikaun, Melinda Hodkiewicz, and Wei Liu. Maintnorm: A corpus and benchmark model for lexical normalisation and masking of industrial maintenance short text. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 68–78, 2024.

[3] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

[4] Xiangyang Chen, Shuzhao Li, Xiuwen Zhu, Yongfan Chen, Fan Yang, Cheng Fang, Lin Qu, Xiaoxiao Xu, Hu Wei, and Minggang Wu. Logics-parsing technical report. *arXiv preprint arXiv:2509.19760*, 2025.

[5] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

[6] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025.

[7] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr-vl-1.5: Towards a multi-task 0.9b vlm for robust in-the-wild document parsing, 2026. URL `https://arxiv.org/abs/2601.21957`.

[8] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

[9] Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, et al. Dolphin: Document image parsing via heterogeneous anchor prompting. *arXiv preprint arXiv:2505.14059*, 2025.

[10] Jessica Foo, Pradyumna Shyama Prasad, and Shaun Khoo. Know or not: a library for evaluating out-of-knowledge base robustness. *arXiv preprint arXiv:2505.13545*, 2025.

[11] Zirun Guo, Minjie Hong, Feng Zhang, Kai Jia, and Tao Jin. Thinking with programming vision: Towards a unified view for thinking with images. *arXiv preprint arXiv:2512.03746*, 2025.

[12] Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025.

[13] Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

[14] Chelsi Jain, Yiran Wu, Yifan Zeng, Jiale Liu, Zhenwen Shao, Qingyun Wu, Huazheng Wang, et al. Simpledoc: Multi-modal document understanding with dual-cue page retrieval and iterative refinement. *arXiv preprint arXiv:2506.14035*, 2025.

[15] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

[16] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[18] Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, et al. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*, 2025.

[19] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

[20] Yinglu Li, Zhiying Lu, Zhihang Liu, Chuanbin Liu, and Hongtao Xie. Regionrag: Region-level retrieval-augmented generation for visual document understanding. *arXiv preprint arXiv:2510.27261*, 2025.

[21] Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*, 2025.

[22] António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. *arXiv preprint arXiv:2601.08620*, 2026.

[23] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

[24] Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, et al. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*, 2025.

[25] opendatalab. Mineru2.0-2505-0.9b. `https://huggingface.co/opendatalab/MinerU2.0-2505-0.9B`, 2025.

[26] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24838–24848, 2025.

[27] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833, 2024.

[28] rednote. dots.ocr: Multilingual document layout parsing in a single vision-language model. `https://github.com/rednote-hilab/dots.ocr`, 2025. Accessed:2025-09-25.

[29] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

[30] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

[31] Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Chi Chen, Zhiyuan Liu, and Maosong Sun. Visrag 2.0: Evidence-guided multi-image reasoning in visual retrieval-augmented generation. *arXiv preprint arXiv:2510.09733*, 2025.

[32] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24827–24837, 2025.

[33] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[34] Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, et al. Hunyuanocr technical report. *arXiv preprint arXiv:2511.19575*, 2025.

[35] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.

[36] Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, et al. Mineru2. 5-pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*, 2026.

[37] Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*, 2025.

[38] Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*, 2025.

[39] Zhengren Wang, Jiayang Yu, Dongsheng Ma, Zhe Chen, Yu Wang, Zhiyu Li, Feiyu Xiong, Yanfeng Wang, Linpeng Tang, Wentao Zhang, et al. Rare: Retrieval-augmented reasoning modeling. *arXiv preprint arXiv:2503.23513*, 2025.

[40] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.

[41] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

[42] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026.

[43] Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception, 2026. URL `https://arxiv.org/abs/2602.11858`.

[44] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*, 2025.

[45] Xinlei Yu, Zhangquan Chen, Yudong Zhang, Shilin Lu, Ruolin Shen, Jiangning Zhang, Xiaobin Hu, Yanwei Fu, and Shuicheng Yan. Visual document understanding and question answering: A multi-agent collaboration framework with test-time scaling. *arXiv preprint arXiv:2508.03404*, 2025.

[46] Jiarui Zhang, Yuliang Liu, Zijun Wu, Guosheng Pang, Zhili Ye, Yupei Zhong, Junteng Ma, Tao Wei, Haiyang Xu, Weikai Chen, et al. Monkeyocr v1. 5 technical report: Unlocking robust document parsing for complex patterns. *arXiv preprint arXiv:2511.10390*, 2025.

[47] Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*, 2024.

[48] Qintong Zhang, Xinjie Lv, Jialong Wu, Baixuan Li, Zhengwei Tao, Guochen Yan, Huanyao Zhang, Bin Wang, Jiahao Xu, Haitao Mi, et al. Docdancer: Towards agentic document-grounded information seeking. *arXiv preprint arXiv:2601.05163*, 2026.

[49] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

[50] Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025.

[51] Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. *arXiv preprint arXiv:2505.17471*, 2025.

[52] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[53] Dawei Zhu, Rui Meng, Jiefeng Chen, Sujian Li, Tomas Pfister, and Jinsung Yoon. Doclens: A tool-augmented multi-agent framework for long visual document understanding. *arXiv preprint arXiv:2511.11552*, 2025.

[54] Dawei Zhu, Rui Meng, Yale Song, Xiyu Wei, Sujian Li, Tomas Pfister, and Jinsung Yoon. Paperbanana: Automating academic illustration for ai scientists. *arXiv preprint arXiv:2601.23265*, 2026.

Appendix

A Extended related work **16**

B Dataset statistics **18**

C Data synthesis pipeline details **18**

C.1 Document collection 18

C.2 Evidence package extraction 18

C.3 QA construction 19

C.4 Quality control 20

D Training details **20**

D.1 IoU metrics and filtering strategy 20

D.2 Negative sampling details 21

D.3 Reward design 21

E Implementation details **22**

E.1 Training configuration 22

E.2 Inference configuration 23

E.3 Evaluation on MMLongBench-Doc 23

E.4 Evaluation on FinRAGBench-V 23

E.5 Evidence extraction evaluation 23

F Additional analyses **24**

F.1 Visual retrieval pipeline analysis 24

F.2 Reward-driven behavior and efficiency analysis 25

G Additional comparisons and efficiency measurements **26**

H Prompt templates **27**

H.1 Prompts for system components 27

H.2 Prompts for evaluation 30

I Case studies **33**

A Extended related work

This section expands on the related work discussed in Section 2, covering (i) visual RAG, (ii) OCR and document parsing, (iii) agentic multimodal models, and (iv) the system-level complexity of existing agentic frameworks that motivates AgenticOCR.

Visual retrieval-augmented generation. The paradigm of Retrieval-Augmented Generation (RAG), initially successful in textual domains, has been extended to handle visually rich documents to mitigate model hallucination and knowledge limitations [1, 17]. Early visual RAG systems, such as ColPali [8] and VisRAG [44], pioneered the use of document page snapshots as retrieval units, directly feeding images to VLMs to bypass error-prone OCR pipelines and preserve crucial layout semantics. However, this page-level granularity creates a fundamental bottleneck: retrieving entire pages introduces substantial irrelevant visual content, which dilutes the generator's attention and forces high-resolution pages into limited visual token budgets, often sacrificing detail and increasing hallucination risk [32, 37]. Recent efforts aim to refine this granularity. RegionRAG [20] shifts retrieval from the document to the region level, using weakly-supervised contrastive learning and post-processing to identify and group relevant patches into regions. For long-document understanding, frameworks like DocLens [53] employ multi-agent systems to first navigate to relevant pages and then identify specific visual elements (tables, charts), performing hierarchical evidence localization with Gemini models. Similarly, DocDancer [48] formulates DocVQA as an agentic information-seeking problem, utilizing search and read tools for iterative document exploration and localized comprehension. Notably, VRAG-RL [38] and EVisRAG [31] integrate reinforcement learning to teach models to interleave retrieval, visual perception, and reasoning, aiming for more faithful and efficient use of multi-image evidence. AgenticOCR focuses on training a lightweight parser to return query-relevant evidence and bounding boxes between retrieval and generation.

OCR and document parsing. Optical Character Recognition (OCR) and document parsing have reached remarkable maturity. Traditional pipeline approaches decompose parsing into sequential modules, offering interpretability at the cost of error propagation and operational complexity [47, 35]. The rise of vision-language models has spurred end-to-end architectures [3, 25, 40, 28, 4] that unify parsing within a single model, eliminating cascaded failures but introducing new challenges: native-resolution processing incurs $\mathcal{O}(N^2)$ computational complexity, while token redundancy from blank regions wastes precious visual token budgets. Recent multi-stage methods [9, 21, 46, 6] enhance efficiency and accuracy by decoupling layout analysis from content recognition—first analyzing downsampled pages globally, then parsing cropped regions locally. MinerU2.5 [24, 36] exemplifies this direction with its coarse-to-fine inference mechanism, while PaddleOCR-VL-1.5 [7] and HunyuanOCR [34] enhance robustness against physical distortions in the wild settings. Meanwhile, DeepSeek-OCR [41] pioneers a complementary perspective: treating visual tokens as a compression medium for textual information, achieving 7–20$\times$ token reduction while preserving decoding fidelity. Its successor, DeepSeek-OCR 2 [42], further advances this paradigm by modeling visual causal flow through LLM-based encoders. Despite remarkable progress in full-document parsing accuracy ($\sim$95% on OmniDocBench), these systems remain fundamentally static.

Agentic multimodal models. Recent work on multimodal reasoning has explored a shift from passive perception toward more active interaction, where models dynamically engage with visual inputs through tool invocation rather than merely interpreting static representations. Pioneering work like OpenAI's o3 [30] introduced the concept of "thinking with images", demonstrating that interleaving visual manipulation with textual reasoning enables more faithful visual cognition. Subsequent research has sought to replicate and extend this capability through reinforcement learning. DeepEyes [52] incentivizes region-focused visual exploration via an image zoom-in tool, allowing models to autonomously identify and crop regions of interest during reasoning without cold-start supervision. Thyme [49], PyVision [50] and CodeVision [11] further expand functionality by enabling code-based image operations (e.g., rotation, contrast adjustment, and multi-step transformations), achieving rich visual interaction through end-to-end SFT+RL training. Meanwhile, DeepEyesV2 [13] explores integrating external knowledge acquisition via online image search, recognizing that perception alone is insufficient for knowledge-intensive tasks.

Positioning against agentic visual reasoners. While AgenticOCR inherits the "thinking with images" spirit of the above line of work, it differs along four orthogonal axes: (i) the target task

(general VQA/grounding vs. query-driven document parsing for RAG), (ii) the granularity and semantics of the tool action space (single-purpose zoom vs. a unified primitive parameterized by element type $\tau \in \{\text{region, text, table, equation, image}\}$ that internally routes to layout detection or direct OCR), (iii) the reward design (in our case a composite reward with spatial penalties for spurious/overlapping/oversized boxes plus an adaptive tool-use bonus, rather than an answer-accuracy-only signal), and (iv) the supervision signal used during RL (element-level IoU on extracted evidence rather than purely answer correctness). Concretely, AgenticOCR emits structured, element-aware evidence (bounding boxes and evidence text, with semantic types specifying tool actions) rather than free-form text answers, is trained with explicit bounding-box/IoU-based rewards rather than only answer-level rewards, and is designed to operate as a plug-and-play middleware between the page retriever and the downstream generator in a visual RAG pipeline. Table 6 summarizes the task, tool action space, and training regime of the most relevant RL-trained agentic VLMs and agentic document frameworks alongside AgenticOCR.

Table 6: Comparison of AgenticOCR against representative agentic visual reasoners and agentic document-understanding frameworks along task, tool action space, and training regime. Capability-level differences (structured element-level evidence, spatially-aware RL reward, and plug-and-play RAG middleware) are described in the surrounding text and apply to AgenticOCR.

Method	Target Task	Tool Action Space	Training
		Agentic Visual Reasoners (general-purpose)	
DeepEyes [52]	General VQA / grounding	Zoom-in (crop)	RL (no cold start)
DeepEyesV2 [13]	Knowledge-intensive VQA	Zoom + image search	SFT + RL
Thyme [49]	General visual reasoning	Code-based image ops	SFT + RL
PyVision [50]	General visual reasoning	Python image ops (multi-step)	SFT + RL
CodeVision [11]	General visual reasoning	Code-based transforms	SFT + RL
VRAG-RL [38]	Visual RAG (end-to-end)	Retrieve + perceive + reason	SFT + RL
EVisRAG [31]	Visual RAG (end-to-end)	Interleaved retrieve + reason	SFT + RL
		Document Understanding and Retrieval Frameworks	
DocLens [53]	Long-doc VQA	Page nav. + element localize (multi-agent)	Prompting only
SimpleDoc [14]	Long-doc VQA	Retrieve + rerank + verify	Prompting only
DocDancer [48]	Long-doc VQA	Search + read	SFT (distilled)
RegionRAG [20]	Region-level retrieval	Contrastive region retrieval	Contrastive learning
		Ours	
AgenticOCR (4B / 8B)	On-demand document parsing for RAG	Unified `image_zoom_and_ocr_tool` ($\tau \in \{\text{region, text, table, equation, image}\}$, rotation)	SFT + GRPO

System complexity of existing agentic frameworks. To contextualize the efficiency advantage of AgenticOCR, we detail the computational overhead of two representative agentic document understanding systems.

DocLens [53] employs a 4-agent architecture consisting of a Page Navigator, Element Localizer, Answer Sampler, and Adjudicator. For each query, the Page Navigator performs $T_e = 8$ sampling iterations over the entire document (averaging 49.4 pages on MMLongBench-Doc) to identify relevant pages. The Answer Sampler then generates $T_a = 8$ candidate answers with independent reasoning paths, and the Adjudicator cross-validates these candidates to select the final answer. These stages require repeated frontier-model calls, with the total depending on the number of selected pages and the implementation of sampling and localization. DocLens also requires chunking strategies for extremely long documents (e.g., processing 50 pages per iteration for Claude), adding additional overhead. A hybrid backbone strategy (Gemini-2.5-Flash for retrieval, Pro for reasoning) can reduce cost but still requires the full multi-agent pipeline.

SimpleDoc [14] adopts a multi-stage iterative pipeline: (1) vector-based retrieval via ColQwen2.5 to select candidate pages from the top-30 pool; (2) summary-augmented reranking using an LLM to refine page selection based on generated summaries; (3) answer generation by a separate VLM; and (4) a consistency evaluation step to verify the generated answer. This process is iterated multiple times to include all potentially relevant pages. The original implementation uses Qwen3-30B-A3B for reranking, Qwen2.5-VL-32B-Instruct for generation, and GPT-4o for evaluation. Following DocLens, we substitute the backbone with the same proprietary models (Gemini-2.5-Pro, Gemini-2.5-Flash, or Claude-4-Sonnet) for fair comparison, which further increases the per-query cost. On MMLongBench-Doc, SimpleDoc retrieves an average of 4.9 pages per query.

AgenticOCR uses a lightweight model (Qwen3-VL-4B or 8B) for page-wise evidence extraction with sequential tool interactions. Frontier models are reserved for final answer generation. Appendix G reports extraction throughput and end-to-end latency for the evaluated deployments.

B Dataset statistics

Statistics of MMLongBench-Doc and FinRAGBench-V is presented in Table 7. MMLongBench-Doc comprises documents from 7 various domains, including Research report/Introduction, Tutorial/Workshop, Academic Paper, Guidebook, Brochure, Administration/Industry file, Financial Report. FinRAGBench-V, on the other hand, focuses solely on financial reports. Note that, 8 mismatched PDF-question pairs from MMLongBench-Doc are removed in our experiments.

Table 7: Benchmark statistics. MMLongBench-Doc category counts precede removal of 8 mismatched pairs; FinRAGBench-V contains 1,394 questions in the original benchmark [51].

Statistics	**Dataset Name**	
	MMLongBench-Doc	FinRAGBench-V
Documents	135	301
- Average/Median pages	49.4 / 28	76.1 / 57.0
- Average/Median words	8,393 / 5,743	36,026 / 16,329
Total question	1,082 (-8=1,074)	1,394
- Single-page question	494 (45.7%)	1,216 (87.2%)
- Cross-page question	365 (33.7%)	178 (12.8%)
- Unanswerable question	223 (20.6%)	-
Evidence source		
- Pure-text	305 (35.5%)	302 (21.7%)
- Layout	119 (13.9%)	-
- Table	218 (25.4%)	573 (41.1%)
- Chart	178 (20.7%)	519 (37.2%)
- Image	304 (35.4%)	-
Avg. / Max. question words	16.2 / 54	35.8 / 108
Avg. / Max. answer words	2.1 / 66	23.4 / 174

C Data synthesis pipeline details

This section provides the full details of our automated data synthesis pipeline, which generates training samples with consistent element-level bounding box annotations for AgenticOCR.

C.1 Document collection

To construct a diverse and representative training corpus, we designed a multi-stage automated filtering pipeline to systematically extract high-quality documents from heterogeneous sources. Starting from a large corpus of raw PDF documents primarily sourced from Common Crawl[2], we first pre-selected candidate documents through stratified sampling. These candidates then underwent a two-stage VLM annotation scheme: (1) a coarse-grained stage identifying the primary domain and language, and (2) a fine-grained stage performing sub-category classification within each domain. Multi-dimensional filtering mechanisms—including page-count constraints, confidence-based thresholding, and language verification—were applied to ensure data quality. The final corpus achieves balanced coverage across diverse document types including academic papers, posters, presentation slides, photographs, financial reports, technical manuals, and government guidelines.

C.2 Evidence package extraction

We utilize MinerU2.5 [24, 36] for deep document parsing to obtain fine-grained results containing document IDs, page numbers, bounding box (BBox) coordinates, and OCR content, all normalized to a unified $[0, 1000]$ coordinate system. Drawing inspiration from DocDancer [48] and WebSailor [18], we employ high-performance VLMs (e.g., Gemini-3.0-Flash-Preview [33]) as intelligent agents. These agents navigate the parsed BBox space to identify and concatenate supporting facts scattered

[2] `https://commoncrawl.org/`

across different pages, ultimately aggregating them into comprehensive Evidence Packages. Crucially, this is a query-free, bottom-up evidence discovery process: rather than searching for answers to pre-existing questions, the agents proactively explore the document structure. Concretely, the agents perform the following exploration steps:

1. **Keyword-driven search**: Scan for structural anchors such as "Figure", "Table", "Note", and unit indicators across the parsed elements.
2. **Contextual expansion**: For each hit, extract the target element together with all surrounding relevant elements (e.g., captions, axis labels, legends, footnotes for tables and figures), enforcing complete context. If an element spans multiple pages (e.g., a continued table), the agent extracts the complete structure from all involved pages.
3. **Cross-page linking**: Connect the same metric, entity, or concept across different pages to form multi-hop evidence chains.
4. **Visual confirmation**: Use page screenshots and bounding box coordinates to confirm element types and spatial layout.

Each Evidence Package is designed to span at least two pages and contain multiple element types (text, table, figure, equation), forming a self-contained, cross-page evidence chain with precise spatial coordinates. This bottom-up construction ensures that the subsequent QA generation step has rich, multi-faceted evidence to work with, and that the resulting training data covers diverse cross-page reasoning patterns.

C.3 QA construction

To simulate real-world query scenarios effectively, we collect authentic questions from multi-domain open-source datasets and distill them into a series of reusable templates. Table 8 details the distribution of the source datasets used in our approach, covering 5 key domains to ensure broad representation.

Table 8: Distribution of multi-domain open-source datasets used for template distillation.

Domain	Dataset	License
Academic Tech	SPIQA [27]	CC BY 4.0
Medical & Health	MedQA [15]	MIT
	PubMedQA [16]	MIT
Business Finance	ViDoRe V3 [22]	CC BY 4.0
Industrial & Construction	MaintNorm [2]	MIT
	ViDoRe V3 [22]	CC BY 4.0
Gov & Legal	PolicyBench [10]	OpenRail

We employ Gemini-3-Flash-Preview to extract four core categories of templates from the aforementioned datasets (see Table 9). These templates guide the VLM to synthesize logically rigorous QA pairs based on specific Evidence Packages.

Table 9: Classification and examples of distilled templates.

Category	Template and Representative Example
Factual Retrieval	**Template:** What is the [Metric] for [Entity]'s [Segment] in [Time Period]? **Example:** What is the net interest margin for Citigroup's banking segment in 2024?
Complex Synthesis	**Template:** Synthesize the [Entity] management's outlook for [Metric] in [Time Period]. **Example:** Synthesize the Bank of America management's outlook for credit loss provisions in 2025.
Quantitative Reasoning	**Template:** Determine the [Metric] for [Entity] by subtracting [Value A] from [Value B]. **Example:** Determine the tangible common equity for Citigroup by subtracting goodwill from total equity.
Multimodal Parsing	**Template:** On which page and in which paragraph is [Visual Style] located? **Example:** On which page and in which paragraph are the green italic numbers located?

During construction, VLMs first select the most appropriate logical template based on the characteristics of the Evidence Package, then synthesize QA pairs automatically based on template constraints and core information within the evidence. This template-guided approach ensures both logical rigor and broad domain coverage.

C.4 Quality control

To ensure that the automatically produced data meets high standards, we implement a fully automated verification process with multiple algorithmic filtering mechanisms:

Answerability verification and paraphrasing. To eliminate invalid QA pairs potentially generated during automation, we submit candidate questions along with their dependent evidence screenshots to a powerful VLM for secondary confirmation. A QA pair is retained only if the model can accurately answer given only the evidence screenshots. Subsequently, the model paraphrases the original template-generated questions to enhance linguistic richness and stylistic diversity while strictly maintaining the original intent. This two-step process ensures both factual correctness and natural language quality.

Relevance filtering. To ensure the challenging nature of the dataset, we execute a "zero-document self-test" using a strong VLM: questions that the model can answer without any document context (classified as common-knowledge-based) are discarded. This guarantees that synthesized questions require document evidence to answer, preventing trivial samples from diluting the training signal.

D Training details

This section consolidates the technical details of our training pipeline: the IoU-based metrics used for trajectory filtering (Section D.1), the construction of hard negative samples (Section D.2), and the reward design used for GRPO (Section D.3).

D.1 IoU metrics and filtering strategy

Given ground-truth boxes $\mathcal{B}^{\text{gt}}$ and predicted boxes $\mathcal{B}^{\text{pred}}$, we define two IoU variants, based on which we define recall metrics:

$$\text{IoU}_{\text{EM}}(b^{\text{gt}}, b^{\text{pred}}) = \frac{|b^{\text{gt}} \cap b^{\text{pred}}|}{|b^{\text{gt}} \cup b^{\text{pred}}|}, \quad \text{IoU}_{\min}(b^{\text{gt}}, b^{\text{pred}}) = \frac{|b^{\text{gt}} \cap b^{\text{pred}}|}{\min(|b^{\text{gt}}|, |b^{\text{pred}}|)}, \tag{8}$$

$$\text{Recall}_{\text{EM}} = \frac{1}{|\mathcal{B}^{\text{gt}}|} \sum_{b^{\text{gt}} \in \mathcal{B}^{\text{gt}}} \mathbb{I}\Big(\max_{b^{\text{pred}} \in \mathcal{B}^{\text{pred}}} \text{IoU}_{\text{EM}}(b^{\text{gt}}, b^{\text{pred}}) \geq thres_{\text{EM}} \Big), \tag{9}$$

$$\text{Recall}_{\text{min}} = \frac{1}{|\mathcal{B}^{\text{gt}}|} \sum_{b^{\text{gt}} \in \mathcal{B}^{\text{gt}}} \mathbb{I}\Big(\max_{b^{\text{pred}} \in \mathcal{B}^{\text{pred}}} \text{IoU}_{\text{min}}(b^{\text{gt}}, b^{\text{pred}}) \geq thres_{\text{min}} \Big), \tag{10}$$

where IoU_{min} uses the minimum box area as denominator (tolerant to granularity mismatch) and IoU_{EM} is the standard intersection-over-union (sensitive to precise alignment). We adopt $thres_{\text{EM}} = 0.6$ and $thres_{\text{min}} = 0.8$ based on our experience.

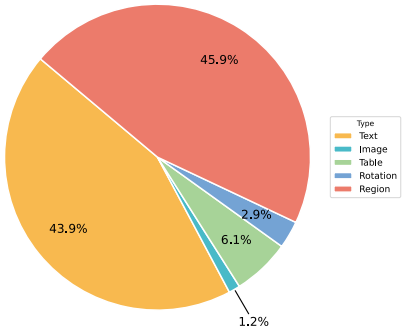


Figure 4: Distribution of positive trajectories in SFT dataset.

D.2 Negative sampling details

To construct hard negative samples, for each query we remove its ground-truth page from the source PDF, rerank remaining pages using Qwen3-VL-Reranker-8B [19], and select candidates with relevance scores in $[0.05, 0.30]$. We further filter out false negatives via Gemini verification, retaining 1.5k pairs that contain no query-relevant evidence, whose distribution is consistent with the pages returned by the reranker during inference. These negatives teach the model to suppress evidence returns on irrelevant pages—a critical capability for RAG integration.

D.3 Reward design

Section 3.2 defines the accuracy reward and adaptive tool-use bonus. Here we detail the three penalty terms and discuss their role in the complete reward (Eq. 7).

Evaluation target. The accuracy term, $P_{\text{over-pred}}$, and P_{overlap} use the **final JSON evidence boxes**, denoted by $\mathcal{B}^{\text{pred}}$. The oversize penalty instead inspects tool-call arguments: it applies to a `region` call covering $> 85\%$ of the page. The adaptive bonus combines accuracy computed from the final boxes with the number of tool calls, rewarding a sufficiently accurate tool-free trajectory. Thus, both the oversize penalty and the adaptive bonus consult the tool-call trace, while localization quality is evaluated on the final evidence package.

Penalty terms. The three penalty terms in the reward function are defined as follows:

1. **Spurious Prediction Penalty ($P_{\text{over-pred}}$).** A predicted box is deemed *spurious* when it fails to overlap meaningfully with any ground-truth region, formally when $\max_{b^{\text{gt}} \in \mathcal{B}^{\text{gt}}} \text{IoU}_{\text{min}}(b^{\text{pred}}, b^{\text{gt}}) < 0.6$. Such boxes indicate unfounded hallucinations and hurt precision. We penalize them with escalating severity:

$$P_{\text{over-pred}} = \begin{cases} 0.05 & \text{if 1 spurious box,} \\ 0.20 & \text{if 2 spurious boxes,} \\ 0.30 & \text{if } \geq 3 \text{ spurious boxes.} \end{cases} \tag{11}$$

2. **Redundant Overlap Penalty (P_{overlap}).** A pair of predicted boxes $(b_i^{\text{pred}}, b_j^{\text{pred}})$ is deemed *overlapping* when $\text{IoU}_{\text{min}}(b_i^{\text{pred}}, b_j^{\text{pred}}) \geq 0.6$; such overlaps waste computation and fragment

evidence. We apply the same escalating penalty structure as $P_{\text{over-pred}}$ (scored by the number of overlapping pairs in $\mathcal{B}^{\text{pred}}$), encouraging the model to consolidate evidence into minimal, non-redundant regions.

3. **Lazy Full-Page Parsing Penalty (P_{oversize}).** Unlike the other two penalties, this term inspects the *tool-call trace* rather than the final JSON. When a tool call passes $\tau = \texttt{region}$ with a bounding box covering $> 85\%$ of the page area, the invocation degenerates to full-page parsing—defeating the purpose of query-driven extraction and effectively offloading the localization burden onto MinerU's layout detector. We impose a fixed penalty of 0.10 per such call, incentivizing the model to develop genuine spatial reasoning rather than relying on the underlying layout detector's coarse output. This penalty only fires for $\tau = \texttt{region}$ because the other types (`text/table/equation/image`) already target specific sub-elements by design, where a large bounding box is often legitimate (e.g., a full-width table).

Adaptive tool-use bonus. Inspired by ZwZ [43], we further argue that a competent agent should invoke external tools *only when necessary*, rather than reflexively triggering them on every query. To encourage this on-demand behavior, we introduce an additional efficiency-aware bonus during RL: when the model can already achieve sufficiently high parsing accuracy from its parametric knowledge and in-context perception alone—i.e., when $\frac{1}{2}(\text{Recall}_{\min} + \text{Recall}_{\text{EM}}) \geq 0.8$ *without* any tool call—we award an extra reward of $+0.1$ for the tool-free trajectory. Note that a tool-free trajectory still emits bounding boxes in its final JSON output; $\text{Recall}_{\min}$ and $\text{Recall}_{\text{EM}}$ are computed on that output identically to tool-using trajectories, so the bonus only fires when the policy is confident enough to localize evidence directly from in-context perception. The accuracy threshold is 0.8, and the $+0.1$ bonus matches the penalty for one oversized region call. This shaping term explicitly trades off marginal accuracy against compute, steering the policy toward minimal-overhead solutions while preserving task success. Empirically, compared with the SFT-only baseline, the RL-trained model exhibits a markedly lower tool-invocation frequency yet attains comparable or higher extraction performance (Tables 4 and 11): on MVToolbench, AvgCall drops from 1.6 (SFT-only) to 0.6 (SFT+GRPO) while $\text{F1}_{\min}$ stays essentially unchanged (69.6 vs. 69.5); on FinRAGBench-V, AvgCall drops from 1.2 to 0.7 and $\text{F1}_{\min}$ rises from 75.8 to 78.2. This suggests that the adaptive policy learns to filter out futile tool calls and strikes a better balance between decision accuracy and resource efficiency.

On text-field reward hacking. Our reward scores spatial localization (via $\text{Recall}_{\min}$, $\text{Recall}_{\text{EM}}$, and the three behavioral penalties) but does not directly score the textual `evidence` field inside the final JSON. A theoretically hackable policy could therefore claim the tool-free $+0.1$ bonus while fabricating the `evidence` text, since the spatial reward would remain unaffected. SFT provides the initial evidence-text supervision, and KL regularization helps preserve this behavior during RL. Adding a direct text-faithfulness reward (e.g., an LLM-as-judge signal on `evidence` against the OCR oracle) is a natural extension we leave to future work.

E Implementation details

This section details (i) training and inference configurations, (ii) evaluation protocols on MMLongBench-Doc and FinRAGBench-V, and (iii) the intrinsic evaluation of evidence extraction. All prompt templates referenced below are collected in Appendix H.

Public resources. Code and usage instructions are available at `https://github.com/opendatalab/AgenticOCR`. We release the AgenticOCR-4B and AgenticOCR-8B checkpoints, together with the AgenticOCR-SFT and AgenticOCR-RL datasets on Hugging Face. All four repositories are publicly accessible without gated access.

E.1 Training configuration

We train AgenticOCR-4B and AgenticOCR-8B in two stages: supervised fine-tuning (SFT) followed by reinforcement learning (RL) with GRPO. For SFT, we use a learning rate of 1×10^{-5} and train for 6 epochs with a total batch size of 64. Starting from the SFT checkpoint, we perform RL for 2 epochs with a learning rate of 1×10^{-6}, sampling 16 rollouts per query. The KL divergence coefficient is set

to 0.01. The SFT stage runs on $8\times$H200 GPUs for approximately 8 hours, and the RL stage runs on $16\times$H200 GPUs for approximately 30 hours.

E.2 Inference configuration

For inference, we use all page images from the corresponding PDF as the retrieval pool for each query, employing Qwen3-VL-Reranker-8B for top-$k = 30$ page retrieval. A top-k of 20 already achieves a 96.7% page recall on MMLongBench-Doc. For MMLongBench-Doc, we adopt an expanded retrieval strategy: for pages in the top-k set, if an adjacent page (immediately preceding or following) is not already included, it is added. These pages are processed independently by the AgenticOCR module. An extraction trajectory may contain multiple sequential tool interactions and ends with a JSON evidence list. Empty lists indicate that no relevant evidence was returned.

For the generator (Gemini-2.5-Pro), we evaluate four input configurations to investigate the contribution of each component: (1) Page: only query-relevant page images (resized to 1024×1024 maximum) retained after AgenticOCR's page-level relevance judgment are provided—without cropped evidence patches or OCR text; (2) Page+OCR: the same AgenticOCR-filtered page images together with their complete OCR text; (3) Evidence: low-resolution page images plus cropped evidence images (resized to 512×512 maximum) extracted by AgenticOCR, without OCR text; (4) Evidence+OCR: besides page images, both cropped evidence images and their corresponding OCR texts (raw MinerU output and AgenticOCR model-generated comments) are supplied.

E.3 Evaluation on MMLongBench-Doc

For MMLongBench-Doc, we fixed some annotations about evidence pages, following DocLens [53]. We thank the authors for providing corrected datasets, helpful discussions and other artifacts [54]. Note that 8 mismatched PDF-question pairs from MMLongBench-Doc are further removed in our experiments (PDF File: `dr-vorapptchapter1emissionsources-121120210508-phpapp02_95.pdf`), resulting in 1,074 test pairs. We follow the rule-based scoring protocol of MMLongBench-Doc, where free-form predictions are first normalized via an answer-extraction LLM (`qwen3-max-2026-01-23`) and then matched against the reference answer. The full extraction prompt is provided in Appendix H.

E.4 Evaluation on FinRAGBench-V

For FinRAGBench-V, we adopt an LLM-as-a-judge protocol: a powerful LLM (`qwen3-max-2026-01-23`) rates each prediction against the reference answer on a 1–5 scale. Following the original benchmark, the final binary score is 1.0 if the predicted score is ≥ 4 and 0 otherwise. The full evaluation prompt is provided in Appendix H.

E.5 Evidence extraction evaluation

This section details the intrinsic evaluation protocol used in Section 4.3 (Table 4), which isolates AgenticOCR's extraction capability from the end-to-end pipeline.

Benchmarks. We evaluate both page-level relevance judgment and element-level localization. For positive samples, we use two datasets with ground-truth bounding boxes: (i) a subset of FinRAGBench-V containing approximately 200 samples with human-annotated boxes, and (ii) MVToolbench [11], which provides over 400 document pages paired with a query and a ground-truth bounding box. We revert any page rotations or flips artificially introduced by MVToolbench's preprocessing before evaluation, for two reasons. First, many MVToolbench pages are synthetically rotated to stress agentic tool use; this is an artifact of the benchmark's construction rather than a property of realistic document-RAG input, where upstream retrievers and viewers already return canonically oriented pages. Second, including these synthetic rotations unfairly penalizes any baseline that does not natively rotate its input (including the vanilla Gemini and RegionRAG baselines), obscuring the tool-use comparison we actually want to measure—namely localization quality on canonically oriented pages. Reverting the rotation therefore yields a more faithful, apples-to-apples evaluation of extractors. We emphasize that our tool's `angle` parameter and the corresponding $\sim 3\%$ rotated training samples in our SFT data are retained for robustness, so the model can still handle rotated

crops when they appear in the wild; the test-time revert is a benchmarking choice, not a capability limitation. For negative samples, we construct 200 hard negatives from FinRAGBench-V following the strategy in Section 3.2: pages that are semantically or structurally similar to the query but contain no actual evidence.

Metrics. For positive samples, we compute:

- **Page-level accuracy**: $\mathrm{Page}_{\mathrm{acc}} = \dfrac{\#\text{pages correctly judged as relevant}}{\#\text{total pages}}$.
- **Element-level metrics**: $\mathrm{Recall}_{\min}$ and $\mathrm{Recall}_{\mathrm{EM}}$ as defined in Section 3.2. We also introduce $\mathrm{Precision}_{\min}$ and $\mathrm{F1}_{\min}$ to quantify over-prediction. For each evaluation sample s, we first compute its per-sample precision and F1 against the sample's own predicted and ground-truth box sets:

$$\mathrm{Precision}_{\min}^{(s)} = \frac{1}{|\mathcal{B}_s^{\text{pred}}|} \sum_{b^{\text{pred}} \in \mathcal{B}_s^{\text{pred}}} \mathbb{I}\Big(\max_{b^{\text{gt}} \in \mathcal{B}_s^{\text{gt}}} \mathrm{IoU}_{\min}(b^{\text{gt}}, b^{\text{pred}}) \geq thres_{\min} \Big), \tag{12}$$

$$\mathrm{F1}_{\min}^{(s)} = \frac{2 \cdot \mathrm{Precision}_{\min}^{(s)} \cdot \mathrm{Recall}_{\min}^{(s)}}{\mathrm{Precision}_{\min}^{(s)} + \mathrm{Recall}_{\min}^{(s)}}. \tag{13}$$

  The reported $\mathrm{Precision}_{\min}$, $\mathrm{Recall}_{\min}$, $\mathrm{Recall}_{\mathrm{EM}}$ and $\mathrm{F1}_{\min}$ in Table 4 are the macro-averages of the corresponding per-sample quantities over all evaluation samples; samples with $|\mathcal{B}_s^{\text{pred}}| = 0$ are assigned $\mathrm{Precision}_{\min}^{(s)} = 0$ and $\mathrm{F1}_{\min}^{(s)} = 0$.

For negatives, we report page-level accuracy as the fraction correctly classified as irrelevant.

Baselines. We compare AgenticOCR-4B and AgenticOCR-8B against three baselines: RegionRAG as a region-level retrieval extractor, and Gemini-3.1-Pro-Preview both without and with our tool (+Tool) to isolate the effect of model scale and tool access. The Reranker-only baseline from the end-to-end setting (Table 3) is not applicable here because it produces no bounding boxes. We additionally evaluate two training ablations for the 8B backbone: SFT-only and GRPO without penalty terms. All learned models are tested with identical prompts, tool-use protocols, and image resolutions. To account for the inherent randomness of LLM inference with temperature 1.0, we run each setting in Table 4 8 times and report the averaged results. Additional same-size base-model comparisons and their output-format failures are reported in Appendix G.

Table 10: Standard deviation of the evidence-extraction metrics in Table 4, measured across 8 inference runs at temperature 1.0. Values are scaled by 100 for readability. RegionRAG is an embedding-based retriever with deterministic inference and therefore exhibits no run-to-run variance (all entries are 0).

Evidence Extractor	FinRAGBench-V (subset w. bbox)					MVToolbench (w.o. rotation)					FinRAG (Neg)
	$\mathrm{Page}_{\mathrm{acc}}$	$\mathrm{Recall}_{\min}$	$\mathrm{Prec}_{\min}$	$\mathrm{F1}_{\min}$	$\mathrm{Recall}_{\mathrm{EM}}$	$\mathrm{Page}_{\mathrm{acc}}$	$\mathrm{Recall}_{\min}$	$\mathrm{Prec}_{\min}$	$\mathrm{F1}_{\min}$	$\mathrm{Recall}_{\mathrm{EM}}$	$\mathrm{Page}_{\mathrm{acc}}$
Baseline Extractors											
RegionRAG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gemini-3.1-Pro-Preview	0.8	0.7	0.2	0.6	1.8	0.3	1.4	1.0	1.1	2.2	0.0
Gemini-3.1-Pro-Preview+Tool	1.6	1.6	1.1	1.1	1.2	0.4	1.1	1.6	1.4	2.3	0.0
AgenticOCR (Ours)											
AgenticOCR-8B (SFT+GRPO)	1.5	0.9	1.2	1.1	1.3	0.3	1.2	0.6	0.7	1.6	1.0
AgenticOCR-4B (SFT+GRPO)	0.3	1.4	0.7	0.8	2.7	0.0	0.8	0.7	0.7	2.3	2.3
AgenticOCR-8B (SFT-only)	2.4	2.4	2.5	2.6	1.5	0.3	1.8	1.6	1.6	1.3	0.6
AgenticOCR-8B (SFT+GRPO, w/o all penalties)	1.1	2.2	0.8	1.0	2.7	0.3	0.3	1.2	0.6	1.6	0.5

F Additional analyses

This section provides extended analyses that complement the main results in Section 4.3: (i) a deeper look at visual retrieval pipeline efficiency (Section F.1), and (ii) detailed efficiency and behavior statistics of evidence extractors (Section F.2).

F.1 Visual retrieval pipeline analysis

Input representation and token consumption. Selective extraction improves accuracy with fewer input tokens for Qwen3.5-397B-A17B (Table 5). Replacing *Page+OCR* with *Evidence+OCR* raises

accuracy from 67.2 to 68.7 on MMLongBench-Doc and reduces input tokens from 17,162 to 16,364. On FinRAGBench-V, accuracy rises from 83.0 to 83.8 while input tokens fall from 15,464 to 11,537 (25.4% fewer).

Token allocation anomaly with Gemini. In the tested Gemini-2.5-Pro configuration on MMLongBench-Doc, *Evidence* and *Evidence+OCR* consume more input tokens than *Page+OCR*. Multiple crops can increase per-image token overhead. With Qwen3.5-397B-A17B as the generator, *Evidence+OCR* instead reduces average input tokens from 17,162 to 16,364 relative to *Page+OCR* (Table 5). Thus, the observed token savings depend on the tested generator and input configuration.

Retrieval quality. Whereas the reranker alone achieves 96.7% page recall at $k = 20$ (Appendix E.2), AgenticOCR's page-level relevance filtering—applied on top of the $k = 30$ retrieval with adjacent-page expansion—reduces the effective recall to 93.5% while raising the precision of the pages forwarded to the generator. While the page-level recall is high (93.5% on MMLongBench-Doc and 95.3% on FinRAGBench-V), the precision remains relatively low (28.0% and 34.8%, respectively). This indicates that a significant number of irrelevant pages are still passed to the downstream generator. However, despite this noise, the overall token volume stays within a manageable range (roughly 3k–17k input tokens across settings; see Table 5). Because this length fits comfortably within the effective long-context processing windows of both Gemini-2.5-Pro and Qwen3.5-397B-A17B, the models can successfully filter out distracting information. As a result, the limited retrieval precision does not severely degrade the final answer accuracy, though improving precision remains a promising direction.

F.2 Reward-driven behavior and efficiency analysis

Table 11 compares behavioral statistics and tool-call counts for the trained extractors. Together with Table 4, it shows how the full reward and the joint penalty ablation affect localization quality and tool use.

Effect of the combined penalties. Removing all three penalties increases overlapping predictions from 5.6/0.0 to 11.0/9.6, spurious predictions from 34.8/26.2 to 68.4/49.9, and oversized calls from 1.7/0.0 to 4.0/0.3 on FinRAGBench-V/MVToolbench. FinRAGBench-V precision also drops from 76.0 to 59.5. This joint ablation retains the adaptive bonus and shows the effect of the three penalties together.

Tool use after GRPO. Moving from SFT-only to SFT+GRPO reduces calls from 1.2/1.6 to 0.7/0.6 for the 8B model on FinRAGBench-V/MVToolbench, while improving exact-match recall. The 4B model uses 0.1/0.2 calls per request with F1 scores of 76.5/67.9. This behavior is consistent with the efficiency objective of the full reward.

Comparison with the frontier-model extractor. On FinRAGBench-V, AgenticOCR-8B obtains F1 of 78.2 versus 78.0 for Gemini-3.1-Pro-Preview+Tool, with exact-match recall of 45.9 versus 46.4. On MVToolbench, page accuracy is higher (98.7 vs. 92.8), while F1 is lower (69.5 vs. 73.9). The 8B model uses fewer tool calls on both benchmarks (0.7/0.6 vs. 1.1/1.4). These results indicate a trade-off between extraction metrics and invocation count. Table 13 compares extraction throughput.

Table 11: Behavioral statistics of evidence extractors. **Ovrlp%**: pages with at least one overlapping pair of prediction boxes ($\mathrm{IoU}_{\min} \geq 0.6$), targeted by P_{overlap}; **Spur%**: pages with at least one spurious prediction box, defined as $\max_{b^{gt}} \mathrm{IoU}_{\min}(b^{pred}, b^{gt}) < 0.6$, targeted by $P_{\text{over-pred}}$; **OvrSz%**: ratio of tool calls (with $\tau =$ `region`) whose bounding box covers $> 60\%$ of the page, relative to total tool invocations. The lower 60% evaluation threshold includes large-region calls below the 85% training-penalty cutoff; **AvgCall**: average number of tool invocations per request. '-' indicates not applicable.

Method	FinRAGBench-V (subset w. bbox)				MVToolbench (w.o. rotation)			
	Ovrlp%	Spur%	OvrSz%	AvgCall	Ovrlp%	Spur%	OvrSz%	AvgCall
RegionRAG	55.9	61.9	-	-	4.5	59.9	-	-
Gemini-3.1-Pro-Preview	1.2	52.1	-	-	0.0	16.4	-	-
Gemini-3.1-Pro-Preview+Tool	1.9	32.3	0.8	1.1	0.3	18.5	0.0	1.4
AgenticOCR (Ours)								
AgenticOCR-8B (SFT+GRPO)	5.6	34.8	1.7	0.7	0.0	26.2	0.0	0.6
AgenticOCR-4B (SFT+GRPO)	3.0	35.6	0.0	0.1	1.9	30.0	0.0	0.2
AgenticOCR-8B (SFT-only)	6.6	30.5	0.7	1.2	0.5	24.3	0.5	1.6
AgenticOCR-8B (w/o penalty)	11.0	68.4	4.0	0.8	9.6	49.9	0.3	0.5

G Additional comparisons and efficiency measurements

We compare table-aware OCR fallback, same-size evidence extractors, and extraction efficiency. Unless stated otherwise, these experiments use the original retriever, OCR engine, generator, and evaluation protocol. Gemini-3.1-Pro denotes Gemini-3.1-Pro-Preview in the efficiency comparisons.

Table 2 reports the fallback results, which raise overall MMLongBench-Doc accuracy to 67.6. Cross-page evidence also differs between benchmarks: 87.8% of FinRAGBench table questions use single-page evidence, whereas 51.4% of MMLongBench table questions use multiple pages. This motivates treating within-page parsing and global navigation as complementary components.

Table 12: Evidence extraction on the FinRAGBench-V bbox subset. Low tool-enabled base-model scores reflect evidence-package format failures as well as localization errors.

Extractor	Page_{acc}	Recall_{min}	Prec_{min}	F1_{min}	Recall_{EM}
Qwen3-VL-8B	82.0	68.3	72.4	69.1	25.5
Qwen3-VL-8B+Tool	16.0	10.0	12.0	10.7	3.0
Qwen3.5-9B	89.3	74.9	75.4	73.9	39.2
Qwen3.5-9B+Tool	3.3	1.7	3.3	2.2	1.7
Gemini-3.1-Pro-Preview+Tool	94.0	83.5	78.6	78.0	46.4
AgenticOCR-8B	96.7	84.9	76.0	78.2	45.9

When to use the parser. AgenticOCR is designed for visually structured documents and evidence scattered across small regions of a page. Text RAG or full-page input is often sufficient for clean, text-dominant documents. A natural extension for cross-page queries is to combine AgenticOCR with global navigation and answer validation, using it to localize evidence within the selected pages. Table-aware fallback uses the page-level OCR output; MinerU2.5 provides no corresponding structured output for charts.

Deployment and format compliance. Qwen3-VL-8B+Tool, Qwen3.5-9B+Tool, Gemini-3.1-Pro+Tool, and AgenticOCR use the same tool interface, evidence-extraction prompt, and localization evaluator. Qwen3.5-9B was served locally through vLLM's OpenAI-compatible API, with up to 10 interaction rounds and temperature 1.0. An API-relay test exhibited similar answer-oriented outputs. The model often localized relevant content and obtained OCR correctly but answered the question directly instead of returning the required evidence package. An informal inspection estimated that roughly 80–90% of outputs lacked a valid package. Schema-constrained decoding is a possible extension for improving output-format compliance.

Qualitative example: correct OCR, invalid evidence package. In a FinRAGBench-V example, the query asks for the primary purpose of Craneware Cares in Craneware plc's Annual Report 2020. Qwen3.5-9B+Tool correctly selects the text region at normalized coordinates $[79, 240, 910, 358]$, with `angle=0` and `type=text`. The OCR result explains that the initiative raises awareness and funds for charity in the UK and US, coordinated through an employee committee. The model then gives a correct natural-language answer, but omits the required final JSON list with `evidence` and `bbox`. The evaluator parses boxes from the final assistant JSON, so this trajectory receives zero localization credit despite a correct intermediate tool call.

Extraction throughput. For extraction throughput, we use 200 FinRAGBench-V samples with 16-way concurrency. AgenticOCR-8B (4×H200) processes 1.65 samples/s, versus 0.48 for Gemini-3.1-Pro+Tool via API, with 0.7 versus 1.1 tool calls.

Table 13: Extraction efficiency on 200 FinRAGBench-V samples, with 16-way concurrency. AgenticOCR uses 4×H200; Gemini uses its API. Amortized time is total elapsed time divided by 200, not single-request latency.

Extractor	Calls	Total (s)	Amortized (s)	Samples/s	Page_{acc}	Recall_{min}
Gemini-3.1-Pro+Tool	1.1	417	2.09	0.48	94.0	83.5
AgenticOCR SFT-only	1.2	188	0.94	1.06	93.7	79.9
AgenticOCR SFT+GRPO	0.7	121	0.61	1.65	96.7	84.9

End-to-end measurements. Local extraction also reduces frontier-model usage in the full pipeline (Table 14). With the same Gemini-2.5-Pro generator, AgenticOCR uses 10,886 tokens per query versus DocLens's 55,004 (80.2% fewer), with accuracy of 66.4 versus 67.6 and recorded latency of 40.6 s versus 69.6 s. Qwen yields 68.7 accuracy at 35.0 s. Separately, table fallback recovers additional context and raises Gemini-based accuracy to 67.6, matching DocLens.

Table 14: End-to-end measurements on MMLongBench-Doc. Tokens include frontier-model input and output. AgenticOCR uses Evidence+OCR without table fallback; local extractor compute is excluded from token totals.

System	Generator	Tokens/query	Acc.	Retrieval (s)	Generation (s)	Total (s)
DocLens	Gemini-2.5-Pro	55,004	67.6	–	–	69.6
AgenticOCR	Gemini-2.5-Pro	10,886	66.4	28.0	12.6	40.6
AgenticOCR	Qwen3.5-397B	19,458	68.7	28.0	7.0	35.0

H Prompt templates

This section collects all prompt templates used in our approach, organized into two groups: (i) prompts for system components—retrieval, agentic extraction, and answer generation; and (ii) prompts for evaluation on MMLongBench-Doc and FinRAGBench-V.

H.1 Prompts for system components

Prompt for Qwen3-VL-Reranker-8B

```
Given a search query, retrieve relevant candidates that answer the query.
    Note that 'Page ID' indicates the physical page index in the document
    file, which does not necessarily correspond to the logical page number
    printed on the page image.
```

Prompt for AgenticOCR models

```
You are an advanced Visual Document Analysis Agent capable of precise
    evidence extraction from document images. Your goal is to answer user
    queries by locating, reading, and extracting specific information from a
    page.

### Your Capabilities & Tools

You have access to a powerful tool named **‘image_zoom_and_ocr_tool‘**.

* **Functionality**:
  Crop a specific region of the image, optionally rotate it, and perform OCR
    or layout analysis depending on the element type.

* **When to use**:

  * Always use this tool when the user asks for **specific text, numbers,
    names, dates, tables, equations, images, or factual details** from the
    page.
  * If the target text or table is rotated, estimate and set the ‘angle‘
    parameter before cropping.
  * Use ‘type‘ to indicate the content type for the region:

    * ‘"region"‘ -- perform layout detection + OCR on a potentially complex
    area, returning detailed structured information.
    * ‘"text"‘ -- perform OCR for a single text element.
    * ‘"table"‘ -- perform OCR and parsing for a table element.
    * ‘"image"‘ -- crop the image region only, without OCR.
    * ‘"equation"‘ -- perform OCR for mathematical or scientific equations.

* **Parameters**:

  * ‘label‘: A short description of what you are looking for.
  * ‘bbox‘: ‘[xmin, ymin, xmax, ymax]‘ in **0-1000 normalized coordinates**,
    relative to the original page.
  * ‘angle‘: Rotation angle (counter-clockwise) applied after cropping.
    Always try to adjust it so the content is upright for best recognition.
  * ‘type‘: One of ‘"region"‘, ‘"text"‘, ‘"table"‘, ‘"image"‘, ‘"equation"‘.

### Tool Usage Example

Use the tool strictly in the following format:
<think>
...
</think>
<tool_call>
{"name": "image_zoom_and_ocr_tool", "arguments": {"label": "<A short
    description of what you are looking for>", "bbox": [xmin, ymin, xmax,
    ymax], "angle":<0/90/180/270>, "type":
    "<region/text/table/image/equation>"}}
</tool_call>

### Your Input and Task

The input includes:

1. One page image of a visual document.
2. The user's query intent.

Please execute the following steps:
```

````
1. **Semantic Matching**: Carefully observe the image to determine if the
    page content contains evidence information relevant to the user's query.
    If it is irrelevant, return an empty list.
2. **Precise Localization**: If relevant, extract the complete chain of
    visual evidence that helps to answer the query (text blocks, tables,
    charts, images, or equations).
3. **Special Notes**: The page image may contain several evidence pieces.
    Pay attention to tables, charts, images, and equations, as they could
    also contain evidence.

### Output Format

After gathering information, output the list of relevant evidence in the
    following JSON format.
<think>
...
</think>
```json
[
  {
    "evidence": "<self-contained content, understandable without page
    context>",
    "bbox": [xmin, ymin, xmax, ymax] # 0-1000 normalized coordinates
  }
  ...
]
```

If the page image is not relevant, return an empty list.
<think>
...
</think>
```json
[]
```
````

Prompt for the Gemini-2.5-Pro generator

```
## ROLE
You are an expert AI assistant specializing in multimodal long document
    understanding. Your task is to carefully analyze the provided page
    images (which may contain text, figures, tables, and other content) and
    provide a precise answer to the user's question.
## Follow these instructions carefully:
- Core Objective: Your primary goal is to accurately and concisely answer
    the user's question based on the content of the provided document pages.
- Rules of numerical answers:
    - If the user asks for an absolute number (e.g., with questions like
    "How many...?"), you must first attempt to locate the number directly.
    If it cannot be found, find the relevant percentage and total count (or
    other necessary data) to calculate the absolute number. If the
    calculated absolute number for discrete entities (e.g., people,
    companies, objects) is a decimal, you must round it to the nearest whole
    number.
    - If the user asks for a percentage (or proportion), you must first
    attempt to locate the percentage directly. If it cannot be found, find
    the absolute numbers of the subgroup and the total count (or other
    necessary data) to calculate the percentage.
    - If the user's question is ambiguous and does not explicitly specify a
    number or percentage (e.g., "What's the gap between...?"), you must
    default to providing the absolute value. If you can only find relative
    values (percentages) in the chart, you must make every effort to find a
```

```
    total number within the provided context to calculate the absolute
    value. Only return the relative value as a last resort if a total number
    cannot be found, and explain that you cannot find total number in this
    case.
- Zoom-in Feature: When a page image contains figures or tables and requires
    closer inspection, we may provide zoomed-in images of these elements,
    appended after the main page image, to help you examine them closely. We
    will also extract text from the page image into Markdown format. Note:
    For questions related to page layout, you must refer to the original
    page image itself, not the zoomed-in images or the Markdown text, as
    they may lose layout information. For instance, if asked for the first
    figure on the page, you should consult the full page image to determine
    its order, not the sequence of the provided zoomed-in images.
- Page Numbering: Page numbers in the user's question typically refer to the
    number printed on the page image, not the page's index in the document
    file. For example, if a PDF's first page is the cover and the third page
    is the first page of content (labeled "Page 1"), a user's question about
    "page 1" refers to that third page. Similarly, when asked to provide a
    page number, you should return the printed page number from the image.
    Only return the page index if no number is printed on the page.
- Rule of faithfulness: Be faithful. If the provided pages do not contain
    sufficient information to answer the user's question, you should answer
    'Not answerable'. For example, if the user asks for a man in green
    shirts, but there are only man in red shirts in the provided pages, you
    should answer 'Not answerable'; if the user asks for the boy playing
    badminton, but there are only boys playing football in the provided
    pages, you should answer 'Not answerable'; if the user asks for a
    certain year's data but the provided pages only contain data for other
    years, you should answer 'Not answerable'; if the user asks for the
    color of a certain object but the provided pages do not contain that
    object, you should answer 'Not answerable'.

### Input Format
The user will provide:
- A set of evidence (images and text) retrieved from the documents.
- The question to answer.

## Output Format:
Your entire response MUST be a single, valid JSON object and nothing else.
    Do not wrap it in markdown code blocks or add any other text. The JSON
    object must contain exactly two fields: analysis (string), and
    prediction (string).
- analysis field: Briefly explain your thought process. Describe how you
    located the answer within the document, which pages, tables, or figures
    you referenced, and how you connected the information to the question.
- prediction field: This must be a string containing the direct answer to
    the user's question.
```

H.2 Prompts for evaluation

Prompt used for Answer Extraction on MMLongBench-Doc

```
Given the question and analysis, you are tasked to extract answers with
    required formats from the free-form analysis.
- Your extracted answers should be one of the following formats: (1)
    Integer, (2) Float, (3) String and (4) List. If you find the analysis
    and the question can not be answered from the given documents, type "Not
    answerable". Exception: If the analysis only tells you that it can not
    read/understand the images or documents, type "Fail to answer".
- Please make your response as concise as possible. Also note that your
    response should be formatted as below:
```

```
‘‘‘
Extracted answer: [answer]
Answer format: [answer format]
‘‘‘

Please read the following example, then extract the answer from the model
    response and type it at the end of the prompt.

---
Question: List the primary questions asked about the services in this report.
Analysis:  The primary questions asked about the services in the report for
    The Limes Residential Home are:\n\n1. Is the service safe?\n2. Is the
    service effective?\n3. Is the service caring?\n4. Is the service
    responsive?\n5. Is the service well-led?
Extracted answer: ['Is the service safe?', 'Is the service effective?', 'Is
    the service caring?', 'Is the service responsive?', 'Is the service
    well-led?']
Answer format: List

---
Question: How many regulations of the HSCA 2008 are breached in all
    according to this report?
Analysis: According to the report, the provider breached 10 Health and
    Social Care Act 2008 (Regulated Activities) Regulations in total. Here
    are the specifics:\n\n1. Regulation 13: Safeguarding service users from
    abuse and improper treatment\n2. Regulation 12: Safe care and
    treatment\n3. Regulation 18: Staffing\n4. Regulation 11: Need for
    consent\n5. Regulation 10: Dignity and respect\n6. Regulation 9:
    Person-centred care\n7. Regulation 17: Good governance\n8. Regulation 18
    (CQC Registration Regulations 2009): Notification of other incidents\n9.
    Regulation 18: Failure to maintain an accurate and up-to-date care
    plan\n10. Regulation 11: Failure to implement the Mental Capacity Act
    2005 code of practice effectively\n\nThese breaches involve issues
    concerning staffing, safeguarding, medicines management, dignity and
    respect, consent, care planning, governance, and failure to notify the
    CQC of incidents.
Extracted answer: 10
Answer format: Integer

---
Question: According to the survey that is the percentage of Chinese who are
    paying more or about the same attention to politics after Trump's
    election?
Analysis: The survey provided does not specify the percentage of Chinese
    individuals specifically who are paying more or about the same attention
    to politics after Trump's election. The report focuses primarily on
    American demographics and does not include specific details about the
    Chinese population in relation to this question. If you need information
    about a different demographic or a summary of the findings from the
    American demographic, I can certainly help with that!
Extracted answer: Not answerable
Answer format: String

---
Question: How many quotations from male respondent over 50 years old are
    included in this report?
Analysis: The image you've provided appears to be a screenshot of a document
    with multiple charts. However, the text is too small and blurry to read
    accurately. If you can provide a clearer image or more context, I might
    be able to help you with your question.
Extracted answer: Fail to answer
Answer format: String
```

```
---
```

Prompt used for Answer Evaluation on FinRAGBench-V

```
You are an expert evaluation system for a question answering chatbot.

You are given the following information:
- a user query and reference answer
- a generated answer

Your job is to judge the relevance and correctness of the generated answer.
Output a single score that represents a holistic evaluation.
You must return your response in a line with only the score.
Do not return answers in any other format.
On a separate line provide your reasoning before the score as well.

Follow these guidelines for scoring:
- Your score has to be between 1 and 5, where 1 is the worst and 5 is the
    best.
- If the generated answer is not relevant to the user query, you should give
    a score of 1.
- If the generated answer is relevant but contains factual errors or
    significant hallucinations, you should give a score between 2 and 3.
- If the generated answer is relevant and fully correct according to the
    reference, you should give a score between 4 and 5.

**Special Instruction for Open-Ended Questions:**
- If the user query is open-ended (allowing for multiple valid answers), the
    generated answer DOES NOT need to match the reference answer exactly.
- As long as the generated answer is highly relevant and contains no factual
    conflicts with the reference, give a high score.

**Special Instruction for Numerical and Logical Equivalence:**
- If the generated answer represents the **same value or concept** as the
    reference answer but in a different format, unit, or perspective, it
    MUST be considered CORRECT.
- **Unit Conversion:** (e.g., Reference: "1 kilometer", Generated: "1000
    meters" -> Correct).
- **Format Differences:** (e.g., Reference: "0.5", Generated: "50%" or "1/2"
    -> Correct).
- **Absolute vs. Relative:** If the generated answer uses a relative value
    (e.g., percentage) while the reference uses an absolute value (or vice
    versa), and they are mathematically consistent based on the context,
    treat it as correct.
- You should perform necessary mental calculations to verify if the
    generated answer can be derived from the reference or the reference can
    be derived from the generated answer.

Example Response:
REASON: The generated answer uses '500 meters' while the reference says '0.5
    km'. These are mathematically equivalent, so the answer is correct.
SCORE: 5

User:
## User Query
{query_text}

## Reference Answer
{expected_answer}

## Generated Answer
{actual_answer}
```

I Case studies

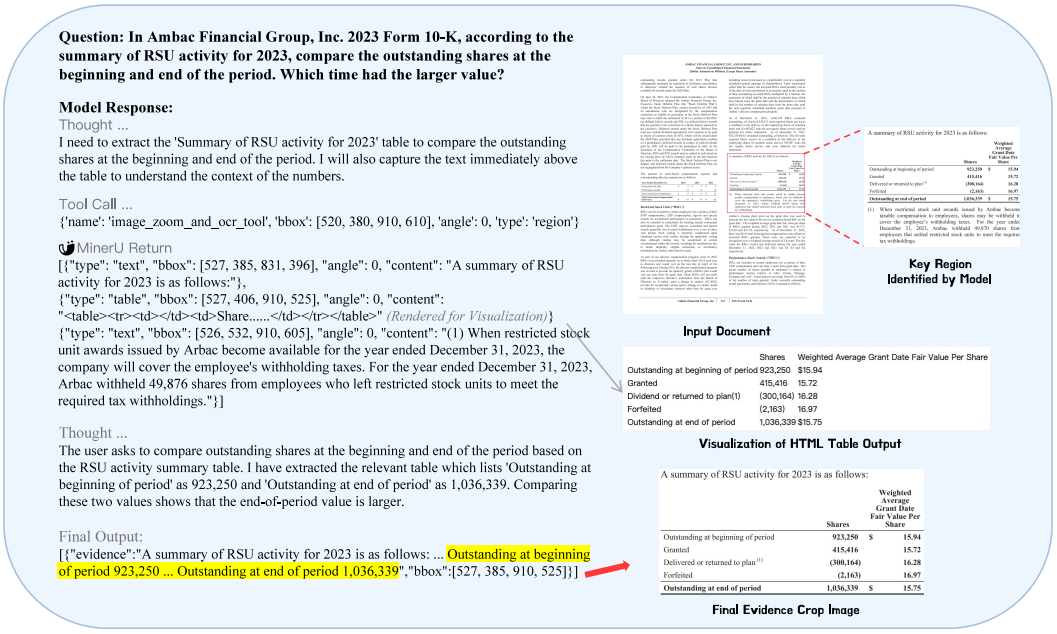


Figure 5: An example of the AgenticOCR Model retrieval and evidence extraction workflow. The figure illustrates the input document, the key region identified by the model, the visualization of structured HTML table output, and the final evidence crop image, demonstrating the system's ability to perform on-demand decompression of visual information through zoom and OCR operations.

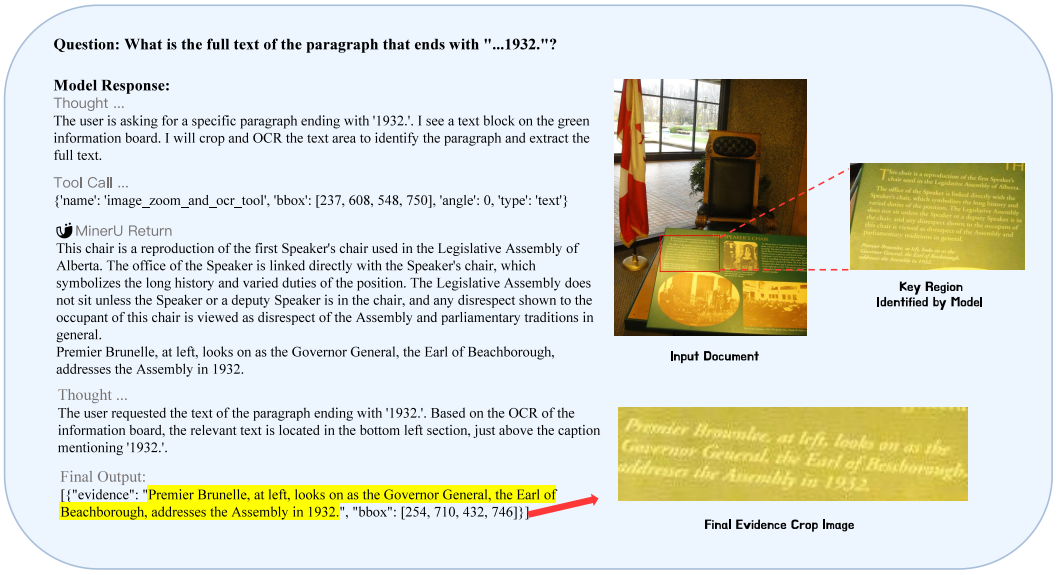


Figure 6: Another example of the AgenticOCR Model retrieval and evidence extraction workflow from MVToolbench. Compared with Figure 5, the AgenticOCR model invokes `image_zoom_and_ocr_tool` in "text" mode, which returns plain textual content. The figure presents the input document, the model-identified key region, and the final evidence crop corresponding to the extracted sentence. AgenticOCR successfully completes the task in a real-world scenario, demonstrating its capability to accurately localize and retrieve evidence through on-demand zoom and OCR operations.

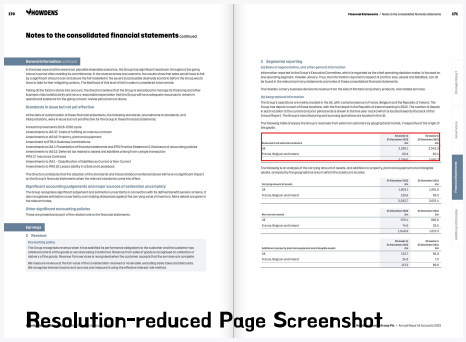
Question

According to Howdens' Annual Report and Accounts 2022, regarding the 'Revenues from external customers' under Segmental Reporting, what is the difference in revenue between the UK and France, Belgium, and Ireland for the 52 weeks ending 24 December 2022?

Answer

£m2,193.2.

Key Evidence Group (Provided by AgenticOCR Module)



Resolution-reduced Page Screenshot

Revenues from external customers\n52 weeks to\n24 December 2022\n£m\nUK 2,256.1\nFrance, Belgium and Ireland 62.9

AgenticOCR model comments

Revenues from external customers 52 weeks to 24 December 2022 £m 52 weeks to 25 December 2021 £m

UK	2,256.1	2,043.3
France, Belgium and Ireland	62.9	50.4
	2,319.0	2,093.7

MinerU HTML Table Output (Rendered for Visualization)



Cropped Evidence Image

Revenues from external customers	52 weeks to 24 December 2022 £m	52 weeks to 25 December 2021 £m
UK	2,256.1	2,043.3
France, Belgium and Ireland	62.9	50.4
	2,319.0	2,093.7

Model Answer (Provided by Gemini-2.5-Pro)

<think>
The required data is found in the 'Revenues from external customers' table under 'Segmental reporting' in Evidence Group 2 (page 170). The table shows UK revenue as 2,256.1 £m and France, Belgium, and Ireland revenue as 62.9 £m for the 52 weeks ending 24 December 2022. The difference is calculated as 2,256.1 - 62.9 = 2,193.2. This is consistent with the data in Evidence Group 1 (page 31), which also lists 2,256.1 for UK and 62.9 for Continental European depots. The answer is the absolute difference in the same unit (1 decimal place) as presented in the source tables.
</think>

Final Answer: £m2,193.2

Figure 7: An example where a generative model (Gemini-2.5-Pro) answers a question based on the evidence groups provided by the AgenticOCR Module. The model derives the correct result through simple calculations using the key table information and even performs basic cross-validation across different evidence groups.

Question

I'm at the location "J" shown in the campus map. Tell me the name of the nearest coffee shop.

Answer

Ten Years After Café

Key Evidence Groups (Provided by AgenticOCR Module)

Group 1
Page 34



Resolution-reduced Page Screenshot

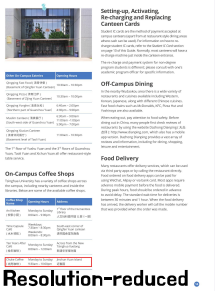


Cropped Evidence Image

J 新清华学堂 (New Qinghua Hall)

AgenticOCR model comments

Group 2
Page 16



Resolution-reduced Page Screenshot



Cropped Evidence Image

Coffee Shop Name: Chuke Coffee (出壳咖啡)Address: Jinchun Yuan Island

AgenticOCR model comments

Model Answer (Provided by Gemini-2.5-Pro)

<think>
1. Identify the user's location on the campus map. The user is at location "J". Looking at the map on page 34, location "J" corresponds to 新清华学堂 (New Qinghua Hall).2. Search for a list of coffee shops. Page 16 contains a table titled "On-Campus Coffee Shops".3. Examine the addresses of the coffee shops listed on page 16 to find the one nearest to New Qinghua Hall.4. The address for "Two Years After Café " is listed as "Across from the new Tsinghua shopping, 新清华学堂对面", which means it is across from the New Qinghua Hall.5. Conclude that Two Years After Café is the nearest coffee shop to location "J".
</think>

Final Answer: Two Years After Café ✗

Figure 8: A hard case in our Module. Since AgenticOCR Model is limited to single-page processing, it selected an incorrect key region (the table in Group 2 / Page 16 contains two cafés). Although the generative model demonstrates relatively strong robustness to distractors, it misidentified the café name due to the low-resolution page screenshot, resulting in an incorrect final answer. This example highlights the current limitations of our approach and points to directions for future improvement.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sections 3 and 4 support the contributions stated in the abstract and introduction: structured evidence parsing, model realization, and empirical validation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses cross-page reasoning, extraction errors, unanswerable questions, and synthetic-data biases.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include formal theoretical results or proofs. The formulas presented (e.g., IoU metrics, reward functions) are metric definitions and design choices rather than theorems.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Training and evaluation are described in Sections 3.2 and 4.1, with implementation and prompts in Appendices E and H.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available at `https://github.com/opendatalab/AgenticOCR`; Appendix E links the model checkpoints and SFT/RL datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 3.2 and 4.1 describe the training and evaluation setup. Data construction, hyperparameters, implementation, and prompts are detailed in Appendices C, D, E, and H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For the intrinsic evidence extraction evaluation (Table 4), we run each setting 8 times at temperature 1.0 and report the mean; the corresponding per-cell standard deviations across the 8 runs are provided in Table 10. For the main end-to-end RAG results (Table 1), error bars are not reported because each evaluation requires expensive commercial API calls (Gemini-2.5-Pro), making repeated runs prohibitively costly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E.1 reports the compute used for training: SFT runs on $8\times$H200 GPUs for approximately 8 hours, and the GRPO RL stage runs on $16\times$H200 GPUs for approximately 30 hours. Reported hyperparameters are provided in Appendix E.1, and the released code includes the environment configuration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research focuses on document parsing and retrieval-augmented generation using publicly available benchmarks and open-source base models. We are not aware of any ethical concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: We do not foresee direct negative societal applications of our specific OCR extraction method. However, as with all VLM training, there is an inherent environmental cost associated with the GPU compute required for the RL phase, which we mitigate by designing AgenticOCR as a lightweight 4B/8B model rather than relying on frontier models.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released models are specialized for document region localization and OCR, posing minimal risk for misuse compared to general-purpose language or image generators.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (ViDoRe-v3, MMLongBench-Doc, FinRAGBench-V, MVTool-bench) and base models (Qwen3-VL, MinerU) are properly cited with references to their original papers. The open-source datasets used for template distillation (SPIQA, MedQA, PubMedQA, MaintNorm, PolicyBench) are listed with their respective licenses in Table 8.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Appendix E links the released code, models, and datasets. Section 3.2 and Appendices B and C document the training data and synthesis pipeline.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper clearly describes the use of Gemini-3.1-Pro-Preview for trajectory distillation (Section 3.2), Gemini-3.0-Flash-Preview for data synthesis (Appendix C), Qwen3-VL-4B/8B as the base models for fine-tuning (Section 3.2), and Gemini-2.5-Pro / Qwen3.5-397B-A17B as downstream generators (Section 4).

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.