
ARIAL: An Agentic Framework for Document VQA with Precise Answer Localization

Ahmad Mohammadshirazi
Ohio State University
Flairsoft
Columbus, Ohio, US
mohammadshirazi.2@osu.edu

Pinaki Prasad Guha Neogi
Ohio State University
Columbus, Ohio, US
guhaneogi.2@osu.edu

Dheeraj Kulshrestha
Flairsoft
Columbus, Ohio, US
dheeraj@flairsoft.net

Rajiv Ramnath
Ohio State University
Columbus, Ohio, US
ramnath.6@osu.edu

Abstract

Document Visual Question Answering (VQA) requires models to not only extract accurate textual answers but also precisely localize them within document images—a capability critical for interpretability in high-stakes applications. However, existing systems achieve strong textual accuracy while producing unreliable spatial grounding, or sacrifice performance for interpretability. We present ARIAL (Agentic Reasoning for Interpretable Answer Localization), a modular framework that orchestrates specialized tools through an LLM-based planning agent to achieve both precise answer extraction and reliable spatial grounding. ARIAL decomposes Document VQA into structured subtasks: OCR-based text extraction with TrOCR, retrieval-augmented context selection using semantic search, answer generation via fine-tuned Gemma 3-27B, and explicit bounding-box localization through text-to-region alignment. This modular architecture produces transparent reasoning traces, enabling tool-level auditability and independent component optimization. We evaluate ARIAL on four benchmarks—DocVQA, FUNSD, CORD, and SROIE—using both textual accuracy (ANLS) and spatial precision (mAP@IoU 0.50:0.95). ARIAL achieves SoTA results across all datasets: 88.7 ANLS and 50.1 mAP on DocVQA, 90.0 ANLS and 50.3 mAP on FUNSD, 85.5 ANLS and 60.2 mAP on CORD, and 93.1 ANLS on SROIE, surpassing the previous best method (DLA) by +2.8 ANLS and +3.9 mAP on DocVQA. Our work demonstrates how agentic orchestration of specialized tools can simultaneously improve performance and interpretability, providing a pathway toward trustworthy, explainable document AI systems. Code is available at: <https://github.com/ahmad-shirazi/ARIAL>

1 Introduction

Document Visual Question Answering (VQA) requires reasoning over both textual content and visual layout in scanned or digitally rendered documents. Models must not only read and understand diverse formats—forms, receipts, reports—but also locate where answers appear within the document structure.

While recent models such as LayoutLMv3 [14], LayoutLLM [29], and DocLayLLM [25] have improved textual accuracy by combining language with layout features, they often treat localization as a secondary task. Consequently, they may generate plausible answers without clearly identifying their source in the document, making verification difficult. Standard metrics like ANLS [40] capture

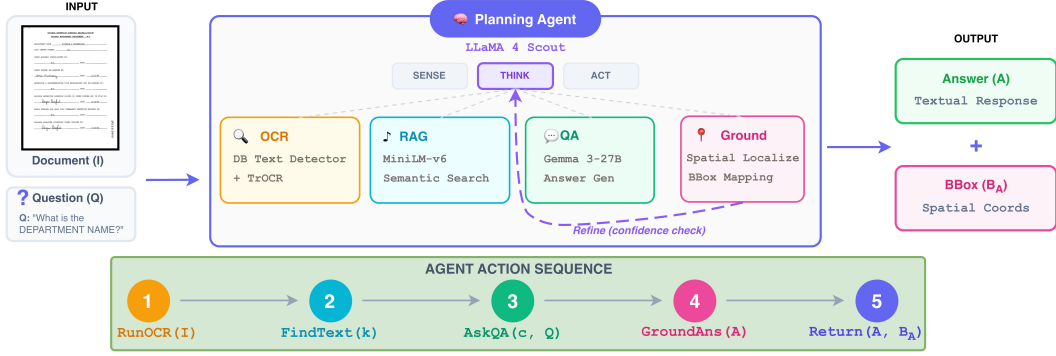


Figure 1: Overview of the ARIAL agentic workflow for Document VQA. The system consists of three modular stages: (1) Input Processing, where an OCR module extracts text segments and bounding boxes from a document image; (2) Agentic Reasoning Pipeline, where the planner agent coordinates task execution—retrieving relevant text, invoking QA or computation, and triggering spatial grounding; and (3) Output Generation, where the final answer and its bounding box are produced. The reasoning loop enables iterative refinement based on confidence, supporting flexible and context-aware decision-making.

string similarity but fail to reflect spatial correctness, prompting a shift towards combined evaluations that include IoU for grounding precision.

DLaVA [33] introduced answer localization by integrating bounding-box prediction within a large multimodal transformer. However, its monolithic design can be computationally intensive and may struggle with fine-grained details in dense or handwritten layouts.

We propose **ARIAL** (Agentic Reasoning for Interpretable Answer Localization), a modular document VQA framework built around an agentic planning model. Rather than using a single large model, ARIAL delegates subtasks—OCR, layout analysis, retrieval, reasoning, and grounding—to specialized modules orchestrated by a central agent. This agent, implemented with LLaMA 4 Scout [31], dynamically selects tools and composes multi-step reasoning chains, enabling accurate and interpretable answers with precise spatial grounding. Our key contributions are:

1. **Agentic Document QA:** We introduce an agent-based document VQA system that decomposes queries into tool calls for OCR, retrieval, and grounding. The modular design enables tool reuse, error tracing, and flexible adaptation across document types.
2. **Precise Answer Localization:** ARIAL produces both answer text and corresponding bounding boxes by aligning answers to OCR-detected spans and contextual cues, ensuring visual traceability.
3. **Retrieval-Augmented Reasoning:** ARIAL incorporates retrieval-augmented generation [19] to focus on relevant text segments, enhancing both reasoning accuracy and efficiency for long or noisy documents.
4. **SoTA Results:** On four benchmarks—DocVQA [30], FUNSD [17], CORD [34], and SROIE [15]—ARIAL achieves new best results in both ANLS and mAP@IoU, reaching 88.7 ANLS and 50.1 mAP on DocVQA.

ARIAL demonstrates how LLMs can be effectively constrained through modular tool orchestration, where each answer is locked to specific pixel coordinates and traceable through interpretable reasoning chains. This addresses fundamental challenges in developing trustworthy, location-aware AI systems for document understanding.

The remainder of this paper is organized as follows: Section 2 reviews related work in document VQA and agentic AI. Section 3 details ARIAL’s architecture and modules. Section 4 outlines datasets and evaluation protocols. Results and analysis appear in Section 5, followed by discussion in Section 6 and conclusions in Section 7.

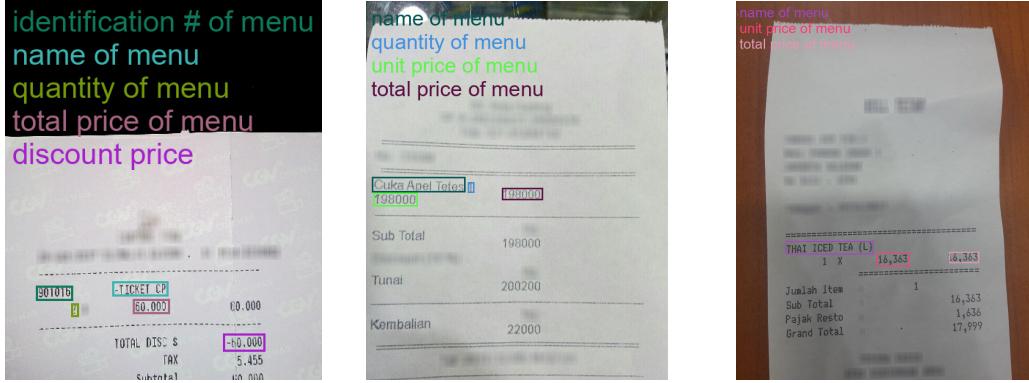


Figure 2: Illustrative examples of visual information extraction on receipt images from the CORD dataset [34]. Each colored annotation corresponds to its extracted answer, highlighted by a matching colored bounding box.

2 Related Work

2.1 Document VQA and Layout-Aware Models

Early document QA systems treated the task as text-only reading comprehension by applying OCR and feeding results into standard NLP models [32]. However, such approaches ignored document structure, prompting the development of layout-aware models. LayoutLM [39], LayoutLMv3 [14], DocFormer [3], and StrucTexT [22] embed both text and spatial coordinates to model document layouts more effectively, achieving strong performance on datasets like DocVQA through unified transformer architectures.

Nevertheless, most models output only answer text and treat localization as auxiliary prediction or post-hoc mapping. Methods like TILT [35] and Donut [18] explore end-to-end generation—Donut bypasses explicit OCR—but lack transparent mechanisms for spatial grounding. As highlighted by DLaVA [33], the inability to visualize answer provenance limits model interpretability and hinders error analysis in high-trust domains.

2.2 Multimodal LLMs for Documents

Multimodal large language models (MLLMs) such as GPT-4o [16], Gemini 2.5 Pro [10], and LLaVA 1.5 [27] extend VQA capabilities by jointly modeling vision and language. These systems answer questions directly from document images using prompt-based interfaces but often function as black boxes, lacking explicit reasoning steps and failing to highlight the visual basis of their answers. Their reliance on global visual understanding can lead to errors in fine-grained text recognition and spatial disambiguation [5].

Recent domain-specific adaptations like LayoutLLM [29] augment prompts with structured spatial cues to guide model focus. DLaVA [33] combines detected text with bounding box metadata or constructed text images, enabling prediction of both answer and spatial location. While DLaVA improves interpretability, it relies on a large, end-to-end multimodal backbone. Our method adopts a modular agentic design enabling more transparent and controllable reasoning while retaining compatibility with any OCR or LLM module.

2.3 Agent-Based and Modular Reasoning

Agentic frameworks have emerged as powerful alternatives to monolithic models for complex tasks [12]. Systems like HuggingGPT [36] use a central language model to coordinate multiple tools for multi-step reasoning. Multi-agent paradigms have been explored for general VQA [41], where specialized agents handle subtasks like reading, counting, or visual interpretation. HAMMR [6] introduces hierarchical architecture improving reasoning granularity and debuggability.

Table 1: Performance comparison on Document VQA datasets using ANLS (textual accuracy).

Category	Method	DocVQA	FUNSD	CORD	SROIE
Text Only	Llama2-7B-Chat [37]	64.99	48.20	47.70	68.97
	Llama3-8B-Instruct [8]	51.79	68.57	52.31	61.24
Text + BBox	LayTextLLM [28]	72.83	78.65	70.81	83.27
Image Only	gpt-oss-20b [1]	79.84	77.64	77.03	80.12
	Llama3.2-11B [8]	78.40	65.02	42.96	61.42
	Pixtral-12B [2]	80.71	78.26	79.08	82.24
	LLaVA-NeXT-13B [26]	51.01	19.71	33.50	13.41
	LLaVA-OneVision-7B [20]	47.59	22.82	32.43	12.10
	Qwen2.5-VL-7B [4]	68.54	58.44	39.01	56.37
	InternVL2-8B [7]	71.26	57.58	55.88	81.55
BBox + Image	DLaVA (Pixtral-12B) [33]	85.9	87.6	84.4	91.4
Text + BBox + Image	LayoutLLM-7B CoT [29]	74.25	78.65	62.21	70.97
	LayoutLLM-7B CoT (Vicuna) [29]	74.27	79.98	63.10	72.12
	DocLayLLM (Llama2-7B) [24]	72.83	78.65	70.81	83.27
	DocLayLLM (Llama3-7B) [24]	78.40	84.12	71.34	84.36
	DLaVA (Pixtral-12B) [33]	74.0	79.6	82.1	91.4
	ARIAL (Ours)	88.7	90.0	85.5	93.1

Table 2: Spatial localization comparison using mAP@IoU (only methods reporting values).

Category	Method	DocVQA	FUNSD	CORD
BBox + Image	DLaVA (Pixtral-12B) [33]	46.2	45.5	57.9
Text + BBox + Image	DLaVA (Pixtral-12B) [33]	34.9	32.0	48.0
	ARIAL (Ours)	50.1	50.3	60.2

In the document domain, MDocAgent [11] employs multiple agents for long-document QA with roles spanning retrieval, modality-specific analysis, key information extraction, and summarization. This modular approach demonstrated notable performance gains, showing the potential of agentic decomposition.

ARIAL builds upon these foundations by tailoring an agentic framework for document VQA. Unlike generic VQA agents, ARIAL handles document-specific challenges such as dense typography, noisy scans, and form-based structures. Its modularity allows independent component upgrades, facilitating efficient domain adaptation and improving interpretability. ARIAL advances document understanding by combining the reasoning power of MLLMs with the transparency and controllability of agentic pipelines, enabling precise answer localization and robust performance across diverse document types.

3 Methodology

3.1 Overview

ARIAL is a modular framework employing a reasoning agent to orchestrate specialized tools for accurate answer generation and precise spatial grounding. The central component is a Planner Agent instantiated by LLaMA 4 Scout, which interprets queries and dynamically routes them through OCR, retrieval, QA, and grounding modules following a sense-think-act paradigm.

Given a document image I and question Q , the system returns answer A and bounding box B_A . The agent constructs a sequence of actions $\{a_1, a_2, \dots, a_n\}$, where each a_i is either a tool call ($\text{RunOCR}(I)$, $\text{FindText}(\text{keywords})$, $\text{AskQA}(\text{context}, Q)$, $\text{GroundAnswer}(\text{answer})$) or an internal reasoning step guiding tool selection. This sequence adapts dynamically to query complexity, terminating when the agent produces a confident answer with visual grounding.

Table 3: Ablation Study (DocVQA and FUNSD)

Model Variant	DocVQA ANLS	DocVQA mAP@IoU	FUNSD ANLS	FUNSD mAP@IoU
<i>Full ARIAL (Agent + RAG + GenQA)</i>	88.7	50.1	90.0	50.3
– No Retrieval (all text to QA)	86.2	48.5	88.1	47.9
– Heuristic Agent (no LLM planning)	83.6	44.2	85.4	42.8
– No Generative QA (lookup only)	87.0	49.0	89.0	49.5

3.2 OCR and Layout Parsing

We employ a two-stage OCR pipeline using DB text detector with ResNet-50 backbone for text region identification, followed by TrOCR for recognition. This yields OCR results $\{(T_i, B_i)\}_{i=1}^N$, where T_i is recognized text and B_i is the corresponding bounding box. Standard preprocessing includes resolution scaling, grayscale conversion, noise removal, and de-skewing. The OCR module maintains reading order and optionally groups segments into structured units using layout heuristics.

3.3 Retrieval-Augmented Generation

The agent performs both lexical and semantic search over OCR segments $\{T_i\}$ using `FindText(keywords)`. Text segments are encoded using MiniLM-v6 Sentence Transformer, with question Q similarly encoded. Retrieved segments $\{(T_j, B_j)\}$ with highest cosine similarity and keyword matches are passed to the QA module. The agent invokes `AskQA(Context, Q)` using Gemma 3-27B [9], which generates answers from retrieved context, reducing hallucination compared to processing entire documents.

For computational queries, the agent identifies relevant numeric fields and invokes `Compute(sum, values)` operations. When no relevant segments are found, the system outputs "No answer found" to avoid unsupported responses.

3.4 Spatial Grounding

After QA generates answer A , the agent invokes `GroundAnswer(A)` to localize the answer. For exact matches to OCR segment T_k , we use bounding box B_k . For multi-segment answers, we merge involved boxes into unified region B_A . For computed answers, the module highlights supporting evidence. Ambiguous answers are disambiguated using contextual cues from retrieved segments and question keywords.

3.5 Training and Fine-Tuning

ARIAL’s modular design enables independent component optimization. OCR uses pretrained DB detector and TrOCR without additional fine-tuning. Retrieval employs off-the-shelf MiniLM-v6 embeddings. The QA module fine-tunes Gemma 3-27B on 70k document QA pairs from DocVQA, CORD, and FUNSD training sets. The Planner Agent uses LLaMA 4 Scout fine-tuned via behavioral cloning on 50 demonstration traces showing appropriate tool usage patterns.

Table 4: End-to-End vs. Agentic Approach Comparison

Metric	LayoutLLM	DocLayLLM	DLA VA OCR-Free	ARIAL (Agentic)
DocVQA ANLS	74.3	78.4	85.9	88.7
DocVQA mAP@IoU	–	–	46.2	50.1
Average Latency (s/q)	0.7	0.4	1.2	3.2
Interpretability	No	No	Yes	Yes + reasoning trace

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate ARIAL on four widely-used document understanding benchmarks:

DocVQA [30] contains 50,000 questions on 12,000+ document images spanning various layouts including forms, receipts, and reports.

FUNSD [17] focuses on form understanding with 9,707 questions across 199 noisy scanned forms, emphasizing spatial relationships and entity linking.

CORD [34] specializes in receipt parsing with 11,000 receipts containing structured fields like menu items, prices, and totals.

SROIE [13] provides 1,000 scanned receipts for information extraction tasks requiring precise key-value pair identification.

We evaluate using two complementary metrics: (1) **ANLS** (Average Normalized Levenshtein Similarity [40]), measuring textual accuracy on a 0–100% scale with tolerance for minor OCR variations, and (2) **mAP@IoU 0.50:0.95**, measuring spatial localization precision by computing mean Average Precision across IoU thresholds from 0.50 to 0.95 in 0.05 increments.

4.2 Baselines and Comparisons

We organize baseline methods into five categories based on their input modalities, as shown in Table 1:

Text Only: Pure language models processing OCR-extracted text without spatial or visual information. We compare against Llama2-7B-Chat [37] and Llama3-8B-Instruct [8], representing strong general-purpose LLMs applied to document text.

Text + BBox: Methods augmenting text with bounding box coordinates. LayTextLLM [28] interleaves layout tokens with text, treating bounding boxes as special tokens within the language model context.

Image Only: Vision-language models processing document images directly without explicit OCR or layout parsing. This category includes:

- gpt-oss-20B [1]: Compact multimodal model optimized for on-device deployment
- Llama3.2-11B [8]: Vision-extended variant of Llama3
- Pixtral-12B [2]: Vision-language model with strong OCR capabilities
- LLaVA-NeXT-13B [26] and LLaVA-OneVision-7B [20]: Advanced visual instruction-tuned models
- Qwen2.5-VL-7B [4]: Recent multimodal model with document understanding focus
- InternVL2-8B [7]: Open-source vision-language model with competitive performance

BBox + Image: Models combining visual features with detected bounding boxes but not explicit text. DLaVA (Pixtral-12B) [33] in OCR-Free mode synthesizes visual text patches, enabling implicit text handling while predicting spatial grounding.

Text + BBox + Image: Methods leveraging all three modalities for comprehensive document understanding:

- LayoutLLM [29]: Instruction-tuned LLM with layout-aware prompting, tested with both base 7B and Vicuna variants using chain-of-thought reasoning
- DocLayLLM [25]: Efficient multimodal extension of LLMs for text-rich documents, evaluated with Llama2-7B and Llama3-7B backbones
- DLaVA (Pixtral-12B) [33]: OCR-Dependent mode using detected text with spatial metadata and image context for answer localization
- ARIAL (Ours): Agentic framework orchestrating specialized tools for OCR, retrieval, reasoning, and spatial grounding

4.3 Implementation Details

Planning Agent: We implement the central orchestration module using LLaMA 4 Scout [31], fine-tuned on 50 curated demonstration traces showing proper tool selection and sequencing patterns. The agent uses 5 in-context few-shot examples for chain-of-thought prompting, enabling dynamic adaptation to query complexity.

OCR Module: Text detection employs the Differentiable Binarization (DB) detector [23] with ResNet-50 backbone, identifying text regions at multiple scales. Recognition uses Microsoft TrOCR [21], a transformer-based OCR engine pretrained on 684M synthetic document images. The pipeline processes pages at approximately 2 seconds per page on NVIDIA H100 GPUs.

Retrieval System: We encode OCR segments using MiniLM-v6 [38], a 384-dimensional sentence transformer optimized for semantic similarity. For efficiency, we retrieve the top-5 most relevant segments for DocVQA and top-3 for FUNSD, CORD, and SROIE, balancing context coverage with computational cost. Retrieval combines dense semantic search (cosine similarity) with sparse keyword matching.

QA Module: The answer generation component uses Gemma 3-27B [9], fine-tuned for 3 epochs on 70,000 document QA pairs sampled from DocVQA, CORD, and FUNSD training sets. We employ the Adam optimizer with learning rate $1e-4$, batch size 16, and gradient accumulation over 4 steps. Training emphasizes generating concise, evidence-grounded answers faithful to retrieved context.

Grounding Module: Spatial localization aligns generated answers to OCR bounding boxes through exact string matching, fuzzy matching (Levenshtein distance ≤ 2), and semantic similarity (≥ 0.85 cosine similarity). For multi-token answers, we compute the union of involved bounding boxes. For numerical computations, we return bounding boxes of all operands.

Infrastructure: Experiments run on $4 \times$ NVIDIA H100 80GB GPUs. The LLaMA 4 Scout agent and Gemma 3-27B QA module are distributed across separate GPUs to enable parallel processing, with the OCR and retrieval modules sharing resources. This configuration achieves an average inference latency of 3.2 seconds per query on DocVQA.

Hyperparameters: We use temperature 0.7 for the planning agent to balance exploration and determinism, and temperature 0.3 for the QA module to prioritize precision. Maximum generation length is set to 128 tokens for answers and 256 tokens for agent reasoning traces. Retrieval cutoff thresholds are 0.5 for semantic similarity and minimum 2 keyword matches for lexical filtering.

5 Results

5.1 Overall Performance

Tables 1 and 2 present our main results. ARIAL consistently achieves SoTA performance on both textual accuracy (ANLS) and spatial localization (mAP@IoU) across all four benchmarks. On DocVQA, ARIAL attains 88.7 ANLS and 50.1 mAP@IoU, representing absolute improvements of +2.8 ANLS and +3.9 mAP points over the previous best method, DLaVA (Pixtral-12B) in OCR-Free mode. On FUNSD, ARIAL achieves 90.0 ANLS and 50.3 mAP@IoU, surpassing DLaVA by +2.4 ANLS and +4.8 mAP points. For receipt datasets CORD and SROIE, ARIAL obtains 85.5 and 93.1 ANLS respectively, with 60.2 mAP@IoU on CORD—outperforming DLaVA by +1.1 ANLS and +2.3 mAP on CORD, and +1.7 ANLS on SROIE.

These consistent improvements across diverse document types—forms, receipts, and general documents—demonstrate the robustness and generalizability of ARIAL’s agentic approach. The simultaneous gains in both textual accuracy and spatial precision highlight the benefit of ARIAL’s modular reasoning and fine-grained retrieval over integrated transformer approaches.

5.2 Comparison Across Input Modalities

Table 1 organizes methods by input modality, revealing important insights about the role of different information sources in document VQA.

Text Only Models demonstrate limited performance, with Llama2-7B-Chat achieving 64.99 ANLS on DocVQA and Llama3-8B-Instruct reaching 51.79 ANLS. These results confirm that pure language

models, despite their strong reasoning capabilities, struggle with document understanding when deprived of spatial and visual context. The particularly poor performance on CORD (47.70 ANLS) and SROIE (68.97 ANLS) suggests that receipt understanding heavily depends on layout cues that text-only approaches cannot capture.

Text + BBox Models like LayTextLLM achieve substantial improvements (72.83 ANLS on DocVQA), demonstrating that explicit spatial coordinates significantly enhance document understanding. The +7.84 point gain over Llama2-7B-Chat shows that layout information is crucial, though still insufficient for SoTA performance.

Image Only Models show highly variable performance. While Pixtral-12B achieves competitive results (80.71 ANLS on DocVQA), other vision-language models struggle significantly. LLaVA-NeXT-13B (51.01 ANLS) and LLaVA-OneVision-7B (47.59 ANLS) perform poorly on DocVQA, suggesting that general-purpose VLMs without document-specific optimization fail to handle dense text and complex layouts. Notably, these models catastrophically fail on receipt datasets (e.g., 12.10 ANLS for LLaVA-OneVision on SROIE), indicating severe limitations in structured document understanding. In contrast, gpt-oss (79.84 ANLS) and InternVL2-8B (71.26 ANLS) demonstrate more robust visual reasoning, though still fall short of multimodal approaches.

BBox + Image Models, represented by DLaVA (Pixtral-12B) in OCR-Free mode, achieve strong performance (85.9 ANLS, 46.2 mAP on DocVQA) by synthesizing visual text patches with predicted bounding boxes. This approach demonstrates that combining visual understanding with spatial grounding yields substantial improvements over image-only methods (+5.19 ANLS over Pixtral-12B baseline).

Text + BBox + Image Models leverage all three modalities for comprehensive understanding. LayoutLLM variants achieve 74.25-74.27 ANLS on DocVQA, while DocLayLLM with Llama3-7B backbone reaches 78.40 ANLS. DLaVA (Pixtral-12B) in OCR-Dependent mode achieves 74.0 ANLS on DocVQA but excels on receipt datasets (82.1 CORD, 91.4 SROIE), showing the value of explicit text integration for structured documents. ARIAL significantly outperforms all methods in this category, achieving 88.7 ANLS on DocVQA—a +10.3 point improvement over DocLayLLM (Llama3-7B) and +14.7 points over LayoutLLM.

5.3 Spatial Localization Performance

Table 2 focuses on spatial grounding capabilities. Only DLaVA and ARIAL report localization metrics, as other baselines do not predict bounding boxes. ARIAL achieves 50.1 mAP@IoU on DocVQA, 50.3 on FUNSD, and 60.2 on CORD, consistently outperforming both DLaVA variants:

- Compared to DLaVA OCR-Free (BBox + Image): ARIAL shows +3.9 mAP on DocVQA, +4.8 mAP on FUNSD, and +2.3 mAP on CORD
- Compared to DLaVA OCR-Dependent (Text + BBox + Image): ARIAL demonstrates even larger gains of +15.2 mAP on DocVQA, +18.3 mAP on FUNSD, and +12.2 mAP on CORD

The substantial mAP improvements reveal that ARIAL’s explicit retrieval-augmented grounding mechanism produces more precise spatial localization than DLaVA’s end-to-end prediction. DLaVA’s OCR-Dependent mode unexpectedly underperforms its OCR-Free mode on spatial grounding (34.9 vs 46.2 mAP on DocVQA), suggesting that integrating explicit OCR text may introduce noise or confusion in its spatial prediction head. In contrast, ARIAL’s modular architecture cleanly separates text understanding from spatial grounding, enabling both superior textual accuracy and precise localization.

5.4 Ablation Study

Table 3 quantifies each component’s contribution to ARIAL’s performance on DocVQA and FUNSD:

No Retrieval: Feeding entire OCR text directly to the QA module causes -2.5 ANLS and -1.6 mAP drops on DocVQA, and -1.9 ANLS and -2.4 mAP drops on FUNSD. These results confirm that targeted context retrieval prevents confusion from irrelevant text and maintains focus on answer-bearing regions. Without retrieval, the QA module must process verbose, noisy OCR output containing hundreds of text segments, leading to attention dilution and increased hallucination risk.

Heuristic Agent: Replacing the LLM-based planning agent with a fixed, rule-based pipeline (always executing RunOCR → FindText → AskQA → GroundAnswer) causes substantial performance degradation: -5.1 ANLS and -5.9 mAP on DocVQA, and -4.6 ANLS and -7.5 mAP on FUNSD. This highlights the value of adaptive, query-aware reasoning. The intelligent agent can recognize when computational queries require arithmetic operations, when answers need multi-hop reasoning across segments, or when retrieval should prioritize semantic versus lexical matches. The heuristic baseline’s inability to adapt leads to systematic errors, particularly on FUNSD’s complex form structures requiring flexible navigation strategies.

No Generative QA: Restricting answer generation to exact string matching from retrieved segments degrades ANLS by -1.7 on DocVQA and -1.0 on FUNSD, while maintaining comparable mAP (-0.1 and -0.8 respectively). This ablation demonstrates the generative QA module’s importance for questions requiring paraphrasing, summarization, or inference beyond direct text spans. For instance, questions like "What is the total cost?" may require summing multiple line items rather than extracting a single value. Spatial grounding remains relatively intact because exact-match answers still align to correct bounding boxes when they exist in the document.

The ablation study confirms that ARIAL’s performance stems from the synergy of all components: intelligent planning, targeted retrieval, generative reasoning, and precise grounding. Removing any component causes measurable degradation, validating the modular design.

6 Discussion

ARIAL’s modular design demonstrates clear advantages over monolithic models through consistent ANLS and mAP gains across diverse document types and structures. The explicit tool orchestration enables both higher textual accuracy (+2.8–10.3 pp over best baselines per dataset) and improved spatial precision (+3.9–18.3 pp in mAP@IoU) compared to prior work.

Interpretability and Trustworthiness: Unlike black-box vision-language models, ARIAL produces transparent reasoning traces showing which tools were invoked, what text segments were retrieved, and how answers were grounded to bounding boxes. This interpretability is crucial for high-stakes applications requiring answer provenance and error diagnosis. When ARIAL produces an incorrect answer, developers can inspect the tool sequence to identify whether the error originated from OCR failure, retrieval miss, QA hallucination, or grounding ambiguity—enabling targeted improvements.

Modularity and Extensibility: ARIAL’s architecture allows independent component upgrades without retraining the entire system. For instance, replacing TrOCR with a more accurate handwriting recognizer would immediately improve performance on handwritten documents. Similarly, incorporating domain-specific QA models (e.g., medical or legal document specialists) requires only swapping the QA module. This modularity facilitates rapid domain adaptation and continuous improvement as better foundation models become available.

Computational Trade-offs: ARIAL incurs higher latency (approximately 3.2 s/query on DocVQA) compared to monolithic models like DocLayLLM (0.4 s) or DLaVA (1.2 s), as shown in Table 4. This overhead stems from sequential tool execution: OCR (2.0s), retrieval (0.3s), QA generation (0.7s), and grounding (0.2s). However, this cost is justified in applications where trustworthiness and explainability are paramount—such as legal document analysis, medical record processing, or financial compliance auditing. For latency-critical applications, ARIAL’s modular design enables optimization through parallelization (e.g., concurrent retrieval and QA) or caching (e.g., reusing OCR results across related queries).

Limitations and Future Work: While ARIAL achieves SoTA results, several limitations warrant attention. The system’s reliance on OCR quality means that documents with severe noise, degradation, or non-standard fonts may produce unreliable results. Second, ARIAL’s sequential processing limits throughput compared to parallelizable end-to-end models. Future work could explore: (1) multi-document reasoning for cross-document QA, (2) active learning to reduce fine-tuning data requirements, and (3) model distillation to compress the agent and QA modules for deployment efficiency.

7 Conclusion

We introduced ARIAL, an agentic framework for Document VQA that emphasizes accurate answer extraction and explicit spatial grounding through modular tool orchestration. By decomposing document understanding into specialized components—OCR, retrieval-augmented generation, answer generation, and spatial localization—coordinated by an LLM-based planning agent, ARIAL achieves state-of-the-art performance across four benchmarks: DocVQA, FUNSD, CORD, and SROIE.

ARIAL surpasses prior methods in both textual accuracy (88.7 ANLS on DocVQA) and spatial precision (50.1 mAP@IoU), demonstrating absolute improvements of +2.8–10.3 ANLS points and +3.9–18.3 mAP points over the strongest baselines per dataset. Our ablation studies confirm that these gains arise from the synergistic combination of intelligent planning, targeted retrieval, generative reasoning, and precise grounding—each contributing measurably to overall performance.

Beyond quantitative metrics, ARIAL’s modular pipeline enables transparent reasoning steps, tool-level auditability, and adaptability to diverse document types—capabilities critically lacking in monolithic models. This makes ARIAL particularly suited for high-stakes settings requiring answer traceability, such as legal document review, medical record analysis, and regulatory compliance monitoring. The explicit separation of concerns allows independent component upgrades and domain-specific customization without full system retraining, facilitating rapid iteration and continuous improvement.

Our work demonstrates the potential of agent-driven AI for document understanding, showing how large language models can be effectively constrained and augmented through explicit tool orchestration rather than unconstrained end-to-end learning. By merging LLM reasoning capabilities with specialized vision and OCR tools under agentic control, ARIAL delivers SoTA performance while meeting real-world demands for trustworthy, explainable, and auditable AI systems. We believe this paradigm—modular, interpretable, and tool-augmented—represents a promising direction for building production-grade document AI that balances performance with transparency.

References

- [1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.
- [2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.
- [3] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Doc-former: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003, 2021.
- [4] Shuai Bai, Kexin Chen, Xiangyu Liu, Jiajie Wang, Weiwei Ge, Sinan Song, Keming Dang, Pei Wang, Shuaipeng Wang, Jiayi Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [5] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.
- [6] Lluís Castrejón, Thomas Mensink, Howard Zhou, Vittorio Ferrari, Andre Araujo, and Jasper Uijlings. Hammr: Hierarchical multimodal react agents for generic vqa. *arXiv preprint arXiv:2404.05465*, 2024.
- [7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

- [9] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL <https://arxiv.org/abs/2503.19786>.
- [10] Google Cloud. Gemini 2.5 pro, June 2025. URL <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro>. Last updated 2025-06-27 UTC.
- [11] 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.
- [12] Kadhim Hayawi and Sakib Shahriar. Ai agents from copilots to coworkers: Historical context, challenges, limitations, implications, and practical guidelines. *Preprints*, 10, 2024.
- [13] Wen Huang, Minghui Qiao, Cong Bai, Yulin Yong, Sheng Zhang, and Qun Guo. Sroie: Scanned receipt ocr and information extraction. In *Proceedings of the ICDAR 2019 Competition on Scanned Receipts OCR and Information Extraction*, 2019.
- [14] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.
- [15] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019.
- [16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [17] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.
- [18] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.
- [19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33. 2020.
- [20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [21] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023.
- [22] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1912–1920, 2021.
- [23] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481, 2020.

- [24] Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*, 2024.
- [25] Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4038–4049, 2025.
- [26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [28] Jilin Lu, Siwen Luo, Srikar Appalaraju, Yusheng Xie, R. Manmatha, and Vijay Mahadevan. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*, 2024.
- [29] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640, 2024.
- [30] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
- [31] Meta. Llama 4 Scout, 2025. URL <https://www.llama.com/docs/get-started/>. Large language model, version released April 5, 2025.
- [32] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.
- [33] Ahmad Mohammadshirazi, Pinaki Prasad Guha Neogi, Ser-Nam Lim, and Rajiv Ramnath. Dlava: Document language and vision assistant for answer localization with enhanced interpretability and trustworthiness. In *Proceedings of the 41st International Conference on Machine Learning*, 2025.
- [34] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.
- [35] Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 732–747. Springer, 2021.
- [36] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.
- [37] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [38] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

- [39] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1192–1200, 2020. doi: 10.1145/3394486.3403172.
- [40] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.
- [41] Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, et al. Igniting language intelligence: The hitchhiker’s guide from chain-of-thought reasoning to language agents. *ACM Computing Surveys*, 57(8): 1–39, 2025.