DOCTREK: TREE-GUIDED AGENTIC REINFORCEMENT LEARNING VIA DENSE REWARDS FOR DOCUMENT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal long-document question answering requires reasoning over hierarchical structures and heterogeneous content across many pages. Existing approaches discard such structures, use the hierarchy only as static retrieval units, or train the policy over flat sequences with trajectory-level feedback. We present **DocTrek**, which couples a section-structured document tree with token-position dense rewards for agentic reinforcement learning. A parsing–construction–integration pipeline converts raw PDFs into a tree whose section skeleton hosts multimodal elements as typed leaves, providing an addressable RL environment. Within it, we train the navigation-and-answering policy with a reward decomposition that converts page-level gold-evidence coverage into rewards placed at action-end tokens, providing dense credit for evidence retrieval and answer generation. Across four benchmarks, DocTrek improves the average score by 2.3 points over the strongest baseline.

1 INTRODUCTION

Long-context documents—such as scientific papers and technical reports—are hierarchical structures interweaving text with figures, tables, and equations. Since human knowledge is predominantly preserved in such formats, automated question answering (QA) over long documents is essential for processing real-world information at scale. Within these documents, information is distributed across multiple pages in diverse modalities, making cross-page integration and cross-modal synthesis the primary challenges.

Existing approaches can be categorized by their processing granularity into three paradigms (Figure 1): text-based RAG Gao et al. (2024), page-as-unit multimodal methods, and element-as-unit multimodal methods. *Text-based RAG* methods retrieve over OCR'd Mishra et al. (2019); Wei et al. (2024) text chunks; however, discarding the visual modality loses information carried by figures, tables, and equations. *Page-as-unit* methods feed rendered page images into multimodal large language models (MLLMs) Bai et al. (2025b); Chen et al. (2025); Liu et al. (2023); however, irrelevant content surrounding the key evidence disperses the model's attention, inducing hallucinations and degrading answer quality. Alternatively, *element-as-unit* methods decompose documents into atomic elements for retrieval. However, this destroys the original layout and structural relationships among elements—a problem known as information fragmentation.

To address these limitations, recent advances have shifted toward recovering structural hierarchies and enabling agentic interaction. Specifically, DMAP Fu et al. (2026) reconstructs document-level structural maps to improve retrieval precision, yet relies on handcrafted, fixed reasoning loops. In parallel, Doc-V^* Zheng et al. (2026) employs reinforcement learning (RL) for sequential page navigation, and DocSeeker Yan et al. (2026) advances agentic RL Feng et al. (2026a;b) through an evidence-aware workflow; however, both operate over a flat sequence of pages, without utilizing document hierarchy. A gap remains: prior agentic RL systems operate over flat page sequences with trajectory-level rewards Lin et al. (2026), without step-level credit anchored on the document hierarchy itself.

We bridge this gap with **DocTrek**, a system that performs tree-guided agentic RL for multimodal long-document understanding. At the foundation of DocTrek lies a *parsing–construction–integration* pipeline that converts raw PDFs into a structured document tree, in which a section-level skeleton

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

hosts text, figure, table, and equation elements as typed leaves. The resulting tree preserves both the section hierarchy and the spatial-semantic links between multimodal elements, providing the environment for downstream agentic learning.

DocTrek optimizes a single orchestration policy end-to-end for document QA. The policy learns when and how to invoke query decomposition, tree navigation, and answer generation, adapting the tool-call sequence to the question and accumulated evidence. To overcome the sparse-reward bottleneck of prior agentic RL, we design a dense process reward that credits each retrieval step by its coverage of ground-truth evidence pages. The tree environment and dense reward are co-designed: the tree exposes addressable states and actions, while the reward supplies page-level evidence feedback at action boundaries, turning hierarchical document structure from a passive retrieval index into an environment for learned navigation.

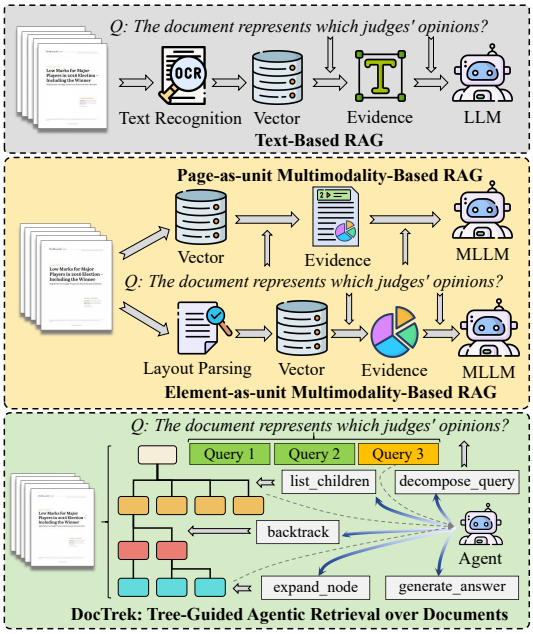

Figure 1: Paradigms for multimodal long-document QA. Text-based RAG (top) drops the visual modality; multimodality-based RAG (middle) keeps it at page or element granularity but still treats the document as a flat unit pool. DocTrek (bottom) represents the document as a hierarchical tree and trains an agent to actively navigate it.

Experiments on four long-document QA benchmarks show that DocTrek outperforms prior methods. Our contributions are threefold: (1) **A Tree-Structured RL Environment for Document QA.** We formulate long-document QA as an MDP (Markov Decision Process) whose state and action spaces are grounded on a section-structured tree with typed multimodal leaves, replacing the flat page sequence with an addressable, explorable structure. (2) **Evidence-Anchored Dense Rewards from Page-Level Supervision.** We design a reward decomposition that converts ground-truth evidence pages into rewards placed at action-end tokens, turning sparse trajectory-level feedback into dense credit for the trajectory prefix ending at each reward-bearing action. (3) **End-to-End Policy Optimization for Document QA.** DocTrek jointly optimizes the main agent's decisions across query decomposition, evidence gathering, and answer generation within a fixed tool environment.

2 Related Work

2.1 Multimodal Long-Document Retrieval and QA

Recent benchmarks Ma et al. (2024); Suri et al. (2025); Dong et al. (2026b) expose the limitations of text-only retrieval over long multimodal documents, motivating a family of methods that retrieve pages or elements with vector-based representations: ColPali Faysse et al. (2025) embeds page screenshots via an MLLM; M3DocRAG Cho et al. (2024) extends multimodal embeddings to closed- and open-domain corpora; and VisDoMRAG Suri et al. (2025) fuses textual and visual RAG under a consistency-constrained modality alignment. MLDocRAG Zhang & Wu (2026) introduces a query–chunk graph that links generated questions to multimodal content. What unites these methods is that they treat the document itself as a flat collection of chunks, pages, or chunk-derived queries, discarding the section hierarchy in which evidence is actually organized.

A subsequent line of work recovers document-intrinsic structure: MoLoRAG Wu et al. (2025) builds a page-level graph for traversal-based retrieval; DocAgent Sun et al. (2025) extracts a tree-formatted outline coupled with an interactive reader; and DMAP Fu et al. (2026) constructs a tri-path structural map driven by reasoning agents. These methods recover structure but do not train a navigation policy over the resulting representation, while agentic RL systems Zheng et al. (2026); Yan et al. (2026); Lin

et al. (2026) learn navigation over flat page sequences. DocTrek connects these directions by using a section-structured tree as the explicit RL environment and placing page-level evidence rewards at action-end positions during navigation.

2.2 Agentic Reinforcement Learning with Process Rewards

Tool-augmented language agents have established the paradigm of autonomous, end-to-end task completion: ReAct Yao et al. (2023) and Toolformer Schick et al. (2023) couple reasoning with tool invocation, removing the modular hand-wiring of earlier pipelines. Subsequent work trains such agents with RL across diverse domains—mathematical reasoning Shao et al. (2024); Guo et al. (2025), multi-turn tool use Cheng et al. (2025); Qian et al. (2025), and web navigation Nakano et al. (2022); Wei et al. (2025)—where the dominant signal is a terminal, outcome-based reward. This terminal-only signal yields sparse gradients in multi-step trajectories: a single scalar credit cannot identify which intermediate decision was responsible for success or failure. Dense process rewards address this bottleneck by supplying step-level supervision: Math-Shepherd Wang et al. (2024) verifies and reinforces each reasoning step in mathematical problem solving; Web-Shepherd Chae et al. (2026) extends step-level process reward models (PRMs) to web navigation with similarly large margins; and parallel work formalizes step-level reward design for general LLM agents Zhang et al. (2025); Choudhury (2025).

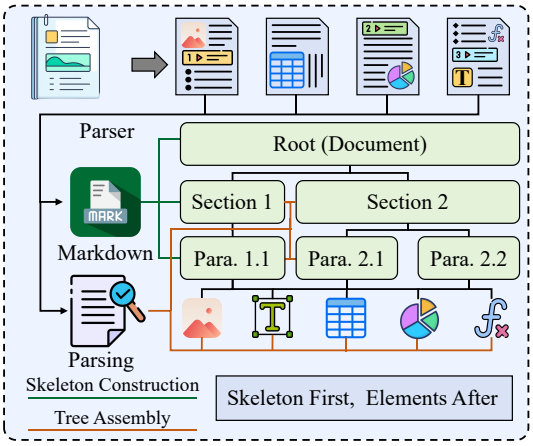

Figure 2: Skeleton-first, elements-after construction of the document tree $\mathcal{T}$. A layout parser produces typed elements. Green edges build the section skeleton from Markdown headings; orange edges then assemble the tree by aligning detected `sec` elements to skeleton nodes and attaching every non-`sec` element as a leaf.

The bottleneck of porting this recipe to document QA is that "what counts as a correct step" must be definable in the target environment. DocTrek closes this gap by treating the document tree as the explicit MDP state space and deriving dense process rewards from ground-truth evidence pages.

3 Hierarchical Document Tree

Documents are hierarchically organized objects in which the section structure determines both the meaning of individual paragraphs and the relationships between evidence elements distributed across pages. To preserve this explicit structure, we propose a hierarchical document tree construction method that follows a skeleton-first, element-later approach, representing the original text and multimodal elements within a unified tree framework.

Tree Formulation. Given a PDF document $\mathcal{D}$ with P pages, we construct a document tree $\mathcal{T} = (V, E)$ with typed nodes V and parent–child edges E, where each node $v \in V$ maps to a specific type $\tau(v) \in \mathcal{T}_{\text{type}} = \{\texttt{root}, \texttt{sec}, \texttt{text}, \texttt{img}, \texttt{tab}, \texttt{eq}\}$. The construction of $\mathcal{T}$ follows a skeleton-first, element-later pipeline across five sequential stages: page structure and layout parsing, element typing, skeleton construction, cross-page element merging, and tree assembly, which we describe below and summarize in Algorithm 1.

Stage 1: Page Structure and Layout Parsing. We invoke an off-the-shelf vision-language document parser f_{parse} that processes each page p_i to output a Markdown-formatted structural parse alongside localized layout elements. Formally, for every page p_i, the parser returns a unified representation:

$$f_{\text{parse}}(p_i) = (\mathcal{L}_i,\ m_i), \tag{1}$$

where m_i is a Markdown-formatted sequence that reconstructs the logical structure, heading depth, and reading order of the page, and $\mathcal{L}_i = \{(b_j, c_j, x_j)\}_j$ is a set of detected elements containing bounding boxes b_j, raw category labels $c_j \in \mathcal{C}_{\text{raw}}$, and content x_j.

Stage 2: Element Typing. We define a mapping ψ to project the 25 raw category labels in $\mathcal{C}_{\text{raw}}$ into either a unified node type set $\mathcal{T}_{\text{type}} = \{\texttt{sec}, \texttt{text}, \texttt{img}, \texttt{tab}, \texttt{eq}\}$ or a discarded set, with the full mapping detailed in Appendix B. For standard text (`sec`, `text`), we directly preserve the extracted plain text. For visual elements (`img`, `tab`, `eq`), we extract their cropped region x_j and invoke an MLLM f_{desc} to obtain a concise natural-language summary $d_j = f_{\text{desc}}(x_j)$. This representation couples visual titles to their counterparts.

Stage 3: Skeleton Construction. We construct the hierarchical skeleton $\mathcal{S}$ of the document tree $\mathcal{T}$ from the Markdown hierarchy. Concatenating $\{m_i\}_{i=1}^{P}$ in page order yields an ordered heading sequence $\langle \sigma_1, \ldots, \sigma_M \rangle$ with associated depths $d(\sigma_m)$. Rooted at a virtual node σ_0 with $d(\sigma_0) = 0$, each σ_m is assigned as a child of its parent

$$\text{parent}(\sigma_m) = \sigma_{m^*}, \qquad m^* = \max \mathcal{A}_m, \tag{2}$$

where $\mathcal{A}_m = \{\, m' \in \{1, \ldots, m-1\} : d(\sigma_{m'}) < d(\sigma_m) \,\}$ is the set of admissible ancestors of σ_m (i.e., headings that appear earlier in reading order and have strictly shallower depth). Each node of $\mathcal{S}$ retains only its heading text from m_i.

Stage 4: Cross-Page Element Merging. Let e_j be the last non-`sec` element of page p_i and e_k the first non-`sec` element of the following page p_{i+1}. We apply a merge operator μ driven by an MLLM-based continuity predicate $\text{Cont}(e_j, e_k) \in \{0, 1\}$:

$$\mu(e_j, e_k) = \begin{cases} e_j \oplus e_k, & \text{Cont}(e_j, e_k) = 1, \\ (e_j, e_k), & \text{otherwise}, \end{cases} \tag{3}$$

where $\oplus$ concatenates textual content and stitches visual elements into a single element. After applying μ at every page boundary, each element fragmented by pagination is restored to a complete semantic unit.

Stage 5: Tree Assembly. The tree $\mathcal{T}$ is assembled from the skeleton $\mathcal{S}$ in two steps. First, each detected `sec` element is aligned to its corresponding skeleton node by heading-text matching, grounding each abstract skeleton node in a concrete, layout-localized section. Second, every non-`sec` element is attached as a leaf: a `text` element e is routed to $\text{sec}(e)$, its nearest preceding section in reading order; a visual element e (`img`, `tab`, `eq`) is routed via its description d_e to the section of its matched title anchor $a(e)$, or to $\text{sec}(e)$ when $a(e) = \varnothing$. The resulting tree preserves both the logical hierarchy and the spatial–semantic links.

4 DocTrek: Tree-Guided Agentic RL

Building upon the hierarchical document tree $\mathcal{T}$, we optimize the retrieval-and-reasoning policy end-to-end over $\mathcal{T}$ via tree-guided agentic RL.

Problem formulation. Given a user question Q and the document tree $\mathcal{T}$, the objective is to generate an answer A by autonomously navigating the tree to gather relevant evidence. We cast this tree-guided document QA process as a finite-horizon Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{R})$ over $\mathcal{T}$.

A state $s_t = (v_t, \mathcal{E}, \mathcal{H}) \in \mathcal{S}$ comprises the current node $v_t \in V$, the accumulated evidence pool $\mathcal{E}$, and the visited-path history $\mathcal{H}$ maintained by the agent. At each step t, the agent perceives the current state and generates a structural operation $a_t \in \mathcal{A}$ following its policy $\pi_\theta(a_t \mid s_t, Q)$. A complete trajectory $o = (s_0, a_0, \ldots, s_T)$ thus traces an explicit navigation path through the tree, terminating with the generation of the final answer A. The policy π_θ is trained to maximize the expected cumulative reward $\mathbb{E}[\mathcal{R}]$.

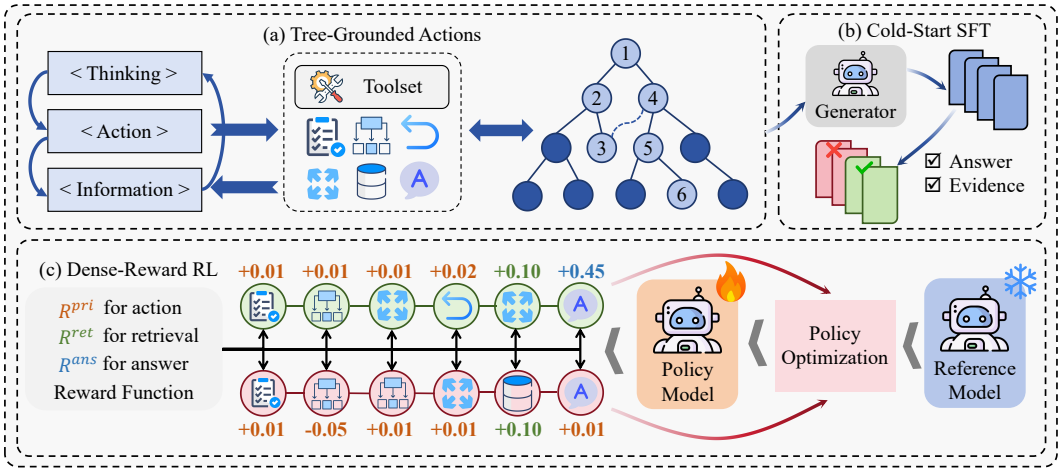

Figure 3: Overview of DocTrek, aligned with §4.1–§4.3. **(a)** The agent interacts with $\mathcal{T}$ via a fixed toolset under the `<thinking>`/`<action>`/`<information>` template, tracing a visited path. **(b)** Cold-start trajectories from a closed-source MLLM are retained for SFT only when both answer and retrieved pages match the gold evidence. **(c)** Each per-turn reward (R^{pri}, R^{ret}) is placed at the last token of its action and the terminal R^{ans} at the final token; GRPO propagates credit backward along the responsible sub-trajectory.

4.1 Tree-Grounded Action Space

To enable the policy $\pi_\theta(a_t \mid s_t, Q)$ to interact with the user question Q and the document tree $\mathcal{T}$, we define the action space $\mathcal{A}$ as a set of structured tools: (1) **`decompose_query`** splits Q into $K \in [1, 3]$ sub-queries targeting distinct tree regions; (2) **`list_children`** returns the children of the current node v_t alongside content previews and their visited status recorded in $\mathcal{H}$; (3) **`expand_node`** fully reads a leaf node to conditionally append it to the active evidence list $\mathcal{E}_k \subseteq \mathcal{E}$ via a relevance judge; (4) **`backtrack`** jumps to an ancestor with unvisited subtrees to recover from dead branches; (5) **`vector_retrieval`** executes a dense top-R retrieval over leaf embeddings; and (6) **`generate_answer`** synthesizes the accumulated evidence $\mathcal{E} = \bigcup_k \mathcal{E}_k$ into the final answer A. Together, these tools support hybrid local–global exploration: `list_children` exposes local breadth over the current node, `backtrack` provides DFS-like recovery, `vector_retrieval` offers global cross-subtree access, and `decompose_query` supports exploration across multiple regions. The execution order is not a hard-coded DFS schedule; the RL objective learns how to interleave tree navigation, dense retrieval, and sub-query re-planning. The full set of prompts used by the agent and the sub-models invoked from tools is listed in Appendix F.

4.2 Cold-Start Initialization

Optimizing π_θ from scratch is brittle: the policy must simultaneously learn the call format, navigation strategy, and answer quality, where a single failure causes degenerate rollouts. We introduce a cold-start stage using a high-quality corpus. We prompt an advanced closed-source MLLM to generate trajectories over $\mathcal{A}$. To prevent the policy from learning to bypass the hierarchical tree, we retain a trajectory for Supervised Fine-Tuning (SFT) only if its final answer is correct and its retrieved pages overlap with the gold evidence. During SFT, the training loss is applied to tokens within the `<thinking>` and `<action>` blocks, while tool-returned `<information>` is masked out. This reason–act–observe template is maintained throughout deployment to prevent format drift.

4.3 RL with Tree-Anchored Dense Rewards

Objective. We optimize π_θ with GRPO Shao et al. (2024), which replaces PPO's Schulman et al. (2017) learned value baseline with a group-relative advantage: for each prompt q, the rollout policy samples G trajectories $\{o_i\}_{i=1}^{G}$, and each advantage is normalized against the in-group statistics. Let $\rho_{i,t} = \pi_\theta(o_{i,t} \mid q, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})$ denote the per-token importance ratio and

$\ell_{i,t} = \min\big(\rho_{i,t}\hat{A}_{i,t},\ \mathrm{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}\big)$ the clipped per-token surrogate. The objective is

$$\mathcal{J}(\theta) = \mathbb{E}_{q,\{o_i\}\sim\pi_{\theta_{\text{old}}}}\Big[\ \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\ell_{i,t} - \ \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)\Big], \tag{4}$$

where π_{ref} is the SFT policy. Standard GRPO assigns a constant advantage $\hat{A}_{i,t}$ uniformly to all tokens in o_i, so terminal feedback alone provides no intermediate temporal anchors in a multi-turn trajectory. Our tree-anchored dense rewards instead place feedback at multiple action-end positions, shortening the propagation distance to preceding decisions while crediting the corresponding trajectory prefixes rather than uniquely identifying a causally responsible action.

Reward instantiation. The reward $\mathcal{R}$ we design is decomposed into three parts that together form the return of a T-turn trajectory o_i:

$$R(o_i) = R^{\text{ans}}(o_i) + \textstyle\sum_{t=1}^{T}\big[R^{\text{ret},(t)} + R^{\text{pri},(t)}\big], \tag{5}$$

where R^{ans} is the terminal reward for answer correctness, augmented by a penalty for an empty $\mathcal{E}$ to prevent tree bypassing. The two per-turn terms are:

$$\begin{aligned} R^{\text{ret},(t)} &= \alpha_t^{\text{hit}}\, r_{\text{hit}}\, N_{\text{hit}}^{(t)}, \\ R^{\text{pri},(t)} &= \alpha_t^{\text{base}} r_{\text{base}} + \alpha_t^{\text{dec}} r_{\text{dec}} + \alpha_t^{\text{back}} r_{\text{back}} + \alpha_t^{\text{fmt}} r_{\text{fmt}}, \end{aligned} \tag{6}$$

where the indicator $\alpha_t^{\bullet} \in \{0,1\}$ activates its term when the corresponding event occurs at turn t, and $N_{\text{hit}}^{(t)}$ counts gold-evidence pages *first* covered at turn t. The five coefficients $r_{\text{hit}}, r_{\text{base}}, r_{\text{dec}}, r_{\text{back}}, r_{\text{fmt}}$ respectively reward evidence hits via `expand_node`, successful tool calls, query decomposition into ≥ 2 sub-queries, backtracking, and penalize format errors. Among them, R^{ret} is the term tied to ground truth, carrying the genuine retrieval signal, while R^{pri} serves as heuristic shaping. A parallel hit reward of the same form is granted to `vector_retrieval` on a separate, smaller budget; we defer its specification to Appendix C. The r_{back} term is unconditional to encourage exploratory recovery. Full triggers and values are in Appendix C.

Token-position placement of rewards. Let $o_i = (x_1, \ldots, x_L)$ denote the token sequence of trajectory o_i, and let $\mathsf{end}_s \in \{1, \ldots, L\}$ be the index of the last token of the s-th turn ($s = 1, \ldots, T$). Let $R^{\text{proc},(s)} = R^{\text{ret},(s)} + R^{\text{pri},(s)}$ denote the per-turn process reward. We deposit each $R^{\text{proc},(s)}$ at position end_s and the terminal reward R^{ans} at position end_T:

$$r_{i,k} = \begin{cases} R^{\text{proc},(T)} + R^{\text{ans}}, & k = \mathsf{end}_T, \\ R^{\text{proc},(s)}, & k = \mathsf{end}_s, \\ 0, & \text{otherwise}. \end{cases} \tag{7}$$

We compute token-level rewards-to-go with $\gamma = 1.0$ and normalize them across the rollout group at each token position, padding missing positions with zeros. Details are provided in Appendix D. A reward deposited at end_s is included without discounting in the returns at or before that position, whereas later returns do not include that earlier reward. Distributing process rewards across action ends therefore provides temporally localized learning signals throughout the trajectory, reinforcing prefixes that reach rewarded intermediate events before the terminal answer.

Two-layer mitigation of reward hacking. Dense process rewards often invite reward hacking Gao et al. (2023); Skalse et al. (2025), where the policy exploits process signals via redundant actions rather than advancing the task. We mitigate this through a two-layer design. At the environment level, the state controller (§4.1) yields no content for revisited nodes in $\mathcal{H}$, preventing R^{ret} from being re-earned. At the reward level, we cap the trajectory-wide process-reward sum at C_{cap} to refine Eq. (5). Letting $S = \sum_{s=1}^{T} R^{\text{proc},(s)}$ be the trajectory-wide process-reward sum, we uniformly rescale each per-turn reward:

$$R^{\text{proc},(s)} \ \leftarrow\ R^{\text{proc},(s)} \cdot \begin{cases} 1, & S \leq C_{\text{cap}}, \\ C_{\text{cap}}/S, & S > C_{\text{cap}}, \end{cases} \qquad C_{\text{cap}} > 0, \tag{8}$$

before placing it at its token position. Once the uncapped sum reaches C_{cap}, additional positively rewarded process actions do not increase the capped process return, preventing redundant tool use from overwhelming the terminal answer signal R^{ans}.

Method	Param.	Paradigm	MMLongBench	LongDocURL	PaperTab	FetaTab	Avg.
LLM-based							
Mistral-7B	7B	RAG	0.245	0.251	0.115	0.411	0.256
Qwen2.5-7B	7B	RAG	0.255	0.279	0.127	0.401	0.266
LLaMA3.1-8B	8B	RAG	0.226	0.298	0.135	0.460	0.280
GPT-4o	-	RAG	0.272	0.327	0.143	0.502	0.311
DeepSeek-V3	671B	RAG	0.298	0.347	0.171	0.524	0.335
MLLM-based							
Qwen2.5-VL-7B	7B	RAG	0.328	0.264	0.298	0.641	0.383
M3DocRAG	7B	RAG	0.362	0.490	0.285	0.638	0.444
MoLoRAG	7B	RAG	0.393	0.517	0.323	0.691	0.481
VisDoMRAG	7B	Agent	-	-	0.299	0.592	-
DocSeeker	7B	Agent	0.401	0.517	-	-	-
Doc-V^*	7B	Agent	0.421	0.563	-	-	-
LAD-RAG	8B	Agent	<u>0.448</u>	0.477	-	-	-
MDocAgent	8B	Agent	0.385	0.469	0.301	0.663	0.455
MAB-DQA	7B	Agent	0.399	0.564	0.269	0.638	0.468
DMAP (GPT-4o)	-	Agent	0.432	**0.607**	<u>0.390</u>	<u>0.725</u>	<u>0.539</u>
DocTrek	8B	Agent	**0.501**	<u>0.574</u>	**0.422**	**0.750**	**0.562**

Table 1: Main results on four multimodal long-document QA benchmarks. Avg. is the mean over the four datasets. Best per column in **bold**, second-best <u>underlined</u>. DMAP's Param. is unknown due to its GPT-4o backbone.

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate on four established multimodal long-document QA benchmarks: **MMLongBench-Doc** Ma et al. (2024), 1,091 questions over 135 PDFs averaging 47.5 pages with cross-page and hallucination probes; **LongDocURL** Deng et al. (2025), targeting understanding, reasoning, and localization over long documents; and the table-centric **PaperTab** and **FetaTab** Hui et al. (2024).

Training Data. Both the cold-start SFT trajectories and the RL rollouts are sampled from two ViDoRe-V3 Loison et al. (2026) subsets: `vidore_v3_pharmaceuticals` and `vidore_v3_hr` for SFT and RL, both disjoint in domain from the four evaluation benchmarks. Detailed statistics and the trajectory filtering procedure are deferred to Appendix D.

Baselines. We compare DocTrek against three baseline paradigms. *(i) LLM-based RAG* performs text-only chunk retrieval over OCR'd content. *(ii) MLLM-based RAG* preserves the visual signal in retrieval through page-image or element-level multimodal embeddings. *(iii) Agent-based* systems navigate structured environments through multi-turn tool use.

Metrics. Following each benchmark's official protocol, for MMLongBench-Doc and LongDocURL we report Generalized Accuracy (Acc), obtained by rule-based evaluation that handles diverse answer formats. For PaperTab and FetaTab we follow MDocAgent Han et al. (2025) and MoLoRAG to employ GPT-4o OpenAI et al. (2024) as evaluator returning Binary Correctness: 1 if the response semantically matches the ground truth, 0 otherwise. The ablation table additionally reports Exact Match (EM) and per-tool average call counts to characterize agent behavior.

Implementation Details. For document processing, we employ Qianfan-OCR Dong et al. (2026a) to perform text recognition and layout parsing; the parser reports a score of 93.12 on OmniDocBench v1.5. We use Qwen3-VL-8B-Instruct for both visual-element description generation and cross-page continuity assessment. The training-free Qwen3-VL-8B-Instruct and Qwen3-VL-8B-Thinking models Bai et al. (2025a) are utilized to execute `decompose_query`, `expand_node`, and `generate_answer`. The `vector_retrieval` tool is powered by Qwen3-VL-Embedding-2B and

Variant	Acc	EM	list_children	expand_node	backtrack	Avg. Tool Call
Training-Free	0.407	0.369	8.93	5.07	5.67	20.69
SFT	0.482	0.432	7.17	7.39	2.32	17.89
w/o Dense Reward	0.488	0.446	8.75	6.96	1.71	18.43
DocTrek	**0.501**	**0.448**	**11.67**	**11.85**	**8.63**	**33.20**

Table 2: Ablation on MMLongBench-Doc. EM = Exact Match. Middle three columns report per-trajectory call counts of the tree-navigating tools; the last column averages over all six tools in $\mathcal{A}$. *Training-Free*: an untrained model used directly as the agent; *SFT*: cold-start only; *w/o Dense Reward*: all rewards placed at the final token.

Qwen3-VL-Reranker-2B Li et al. (2026). The cold-start SFT trajectories are generated by the closed-source Qwen3.5-Plus model. During the training phase, both the policy actor and the inline answer judge for scoring R^{ans} are instantiated from Qwen3-VL-8B-Instruct. RL updates only the main agent; all tool-internal subagents remain frozen. End-to-end optimization refers to policy-gradient learning of orchestration across the complete QA workflow. The reward coefficients listed in Appendix C are fixed across all four evaluation benchmarks, with no benchmark-specific tuning. Tree construction is performed offline once per document at 13.8 s/page. Training uses 24 H20 GPUs for approximately 120 hours (2,880 GPU-hours). On MMLongBench-Doc, inference averages 124 s/query and 33.2 tool calls. Detailed training settings are deferred to Appendix D.

5.2 Main Results

Table 1 compares the end-to-end performance of the three baseline paradigms under each method's published configuration. DocTrek achieves the highest average score (**0.562**) and ranks first on three of the four benchmarks (MMLongBench-Doc, PaperTab, FetaTab), placing second only on LongDocURL—behind DMAP, whose backbone is GPT-4o.

vs. LLM-based RAG. LLM-based RAG discards the visual modality by feeding OCR'd text into a text-only LLM Jiang et al. (2023); Team (2024); Grattafiori et al. (2024). Even DeepSeek-V3 DeepSeek-AI et al. (2025) with 671B parameters reaches only 0.335 average—22.7 points below DocTrek. Parameter scaling does not close the gap because a document's tables, figures, and equations carry information that OCR cannot fully recover. DocTrek instead preserves these signals as first-class typed leaves of its document tree, which is why an 8B multimodal agent reads from richer evidence than any text-only RAG can access, regardless of model size.

vs. MLLM-based RAG. MLLM-based RAG preserves visual content but primarily relies on coarse pages as retrieval units. The strongest baseline in this group, MoLoRAG, reaches 0.481 average—8.1 points below DocTrek, and 10.8 points below on MMLongBench-Doc. Page-level retrieval breaks down whenever the answer requires composing information across separated regions: pieces of a table split by a page boundary, or a figure and its caption interrupted by a page break. Furthermore, feeding entire rendered pages into the model introduces irrelevant surrounding content that disperses attention. DocTrek's hierarchical tree resolves these cases—cross-page elements are merged into single semantic units and the section hierarchy lets the agent isolate the exact evidence without being overwhelmed by whole-page noise.

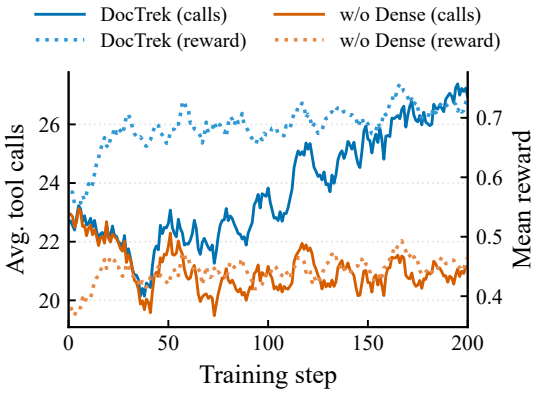

Figure 4: Training dynamics over the first 200 GRPO steps. Average tool calls and mean reward increase for DocTrek, whereas the w/o Dense variant plateaus on both axes.

vs. Agent-based methods. Agent-based methods Sourati et al. (2025); Xiang et al. (2026)

span diverse designs, including evidence-aware workflows (DocSeeker), page-level GRPO (Doc-V^*), and structural maps (DMAP). DocTrek surpasses all 7–8B agents across every benchmark, and notably wins by the largest margin on MMLongBench-Doc (+10.0 over DocSeeker, +8.0 over Doc-V^*), where cross-page evidence integration is most demanded—suggesting that *tree-anchored navigation* is the important factor when evidence is fragmented across the document. Furthermore, against DMAP—which employs the larger GPT-4o backbone—DocTrek still wins on three of the four benchmarks (MMLongBench-Doc **+6.9**, PaperTab **+3.2**, FetaTab **+2.5**) and on the overall average. These results compare complete systems under their published configurations.

5.3 Ablation Studies

To evaluate the individual contribution of each component within DocTrek's training pipeline, we conduct a systematic ablation study on MMLongBench-Doc (Table 2). We begin by replacing the dense process reward with a terminal, outcome-only reward scheme (w/o Dense Reward). Compared with w/o Dense Reward, DocTrek improves Accuracy and Exact Match by 1.3 and 0.2 percentage points, respectively, while increasing average tool calls from 18.43 to 33.20. Recall@10 also increases from 59.72% to 68.85% (Appendix E), indicating improved evidence coverage. Since Recall@10 uses a fixed cutoff, the improvement reflects better gold-evidence coverage within the first ten retrieved leaves. These gains are accompanied by greater inference effort.

Removing the RL phase entirely (SFT) yields a 1.9-point Accuracy drop. The Training-Free setting further reduces Accuracy to 0.407, 9.4 points below DocTrek. This result shows that cold-start trajectories contribute substantially to the agent's final performance. Although the Training-Free baseline issues 5.67 `backtrack` calls per trajectory, this action frequency does not translate into comparable answer accuracy. This delineates a clear division of labor: SFT instills the baseline operational standard, while dense-reward RL elevates this syntactic foundation into purposeful exploration.

Variant	Acc	EM	Avg. Tool Call
w/o VR	0.414	0.371	33.90
w/o CPM	0.486	0.442	37.60
DocTrek	**0.501**	**0.448**	**33.20**

Table 3: Component ablations on MMLongBench-Doc. w/o VR disables vector retrieval; w/o CPM leaves cross-page elements unmerged; DocTrek uses both components. Acc and EM measure answer quality, while Avg. Tool Call reports the mean number of tool invocations per query.

Figure 4 shows average tool calls and mean reward over the first 200 GRPO steps. DocTrek's mean reward rises early and stays above the ablation, whereas its tool-call count increases after an initial decline. The ablation's tool-call count and reward continue to fluctuate around similar levels. Together with the retrieval gains in Appendix E, these patterns are consistent with more extensive evidence gathering under dense feedback.

We further evaluate the contributions of cross-page merging during tree construction and the vector-retrieval tool in Table 3. Removing vector retrieval produces the larger degradation, reducing Accuracy by 8.7 points and EM by 7.7 points. Without cross-page merging, content belonging to the same cross-page element remains split across multiple nodes. The agent therefore requires 4.40 more tool calls on average to search for the dispersed evidence, while Accuracy and EM decrease by 1.5 and 0.6 points, respectively.

6 Conclusion

We presented DocTrek, which couples a section-structured document tree with token-position dense rewards for multimodal long-document QA: the tree provides an addressable RL environment, and a reward decomposition converts gold-evidence overlap into rewards placed at action-end tokens, providing temporally localized credit to trajectory prefixes that reach rewarded intermediate events. This design provides GRPO with temporally localized training signals across multi-turn navigation trajectories. Across four long-document QA benchmarks, DocTrek outperforms prior open-source systems, with ablations showing performance gains from dense reward placement, cross-page merging, and vector retrieval within the proposed system.

AI USE STATEMENT

In this work, we used Qwen3.5-Plus to generate candidate trajectories for supervised fine-tuning, as described in Appendix D. We also used generative AI tools to assist with translation and provide feedback on methodological descriptions and the interpretation of experimental results. Additionally, we used these tools to adjust manuscript formatting and identify relevant literature. The authors reviewed and approved the AI-assisted changes and take responsibility for the final content of this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Hyungjoo Chae, Seonghwan Kim, Junhee Cho, Seungone Kim, Seungjun Moon, Gyeom Hwangbo, Dongha Lim, Minjin Kim, Yeonjun Hwang, Minju Gwak, et al. Web-shepherd: Advancing prms for reinforcing web agents. *Advances in Neural Information Processing Systems*, 38:63314–63356, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Mingyue Cheng, Jie Ouyang, Shuo Yu, Ruiran Yan, Yucong Luo, Zirui Liu, Daoyu Wang, Qi Liu, and Enhong Chen. Agent-r1: Training powerful llm agents with end-to-end reinforcement learning, 2025. URL https://arxiv.org/abs/2511.14460.

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, 2024. URL https://arxiv.org/abs/2411.04952.

Sanjiban Choudhury. Process reward models for llm agents: Practical framework and directions, 2025. URL https://arxiv.org/abs/2502.10325.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan

Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.57. URL https://aclanthology.org/2025.acl-long.57/.

Daxiang Dong, Mingming Zheng, Dong Xu, Chunhua Luo, Bairong Zhuang, Yuxuan Li, Ruoyun He, Haoran Wang, Wenyu Zhang, Wenbo Wang, Yicheng Wang, Xue Xiong, Ayong Zheng, Xiaoying Zuo, Ziwei Ou, Jingnan Gu, Quanhao Guo, Jianmin Wu, Dawei Yin, and Dou Shen. Qianfan-ocr: A unified end-to-end model for document intelligence, 2026a. URL https://arxiv.org/abs/2603.13398.

Kuicai Dong, CHANG YUJING, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimomal generation for document question answering. *Advances in Neural Information Processing Systems*, 38, 2026b.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=tRk1nofSmz.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026b.

Shunliang Fu, Yanxin Zhang, Yixin Xiang, Xiaoyu Du, and Jinhui Tang. DMAP: human-aligned structural document map for multimodal document understanding. In Hakim Hacid, Yoelle Maarek, Francesco Bonchi, Ido Guy, and Emine Yilmaz (eds.), *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pp. 2037–2048. ACM, 2026. doi: 10.1145/3774904.3792229. URL https://doi.org/10.1145/3774904.3792229.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and

Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily

Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia

Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

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, 2025. URL https://arxiv.org/abs/2503.13964.

Yulong Hui, Yao Lu, and Huanchen Zhang. UDA: A benchmark suite for retrieval augmented generation in real-world document analysis. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7c06759d1a8567f087b02e8589454917-Abstract-Datasets_and_Benchmarks_Track.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL https://arxiv.org/abs/2601.04720.

Jiahang Lin, Kai Hu, Binghai Wang, Yuhao Zhou, Zhiheng Xi, Honglin Guo, Shichun Liu, Junzhe Wang, Shihan Dou, Enyu Zhou, Hang Yan, Zhenhua Han, Tao Gui, Qi Zhang, and Xuanjing Huang. Mm-doc-r1: Training agents for long document visual question answering through multi-turn reinforcement learning, 2026. URL https://arxiv.org/abs/2604.13579.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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. *CoRR*, abs/2601.08620, 2026. doi: 10.48550/ARXIV.2601.08620. URL https://doi.org/10.48550/arXiv.2601.08620.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ae0e43289bffea0c1fa34633fc608e92-Abstract-Datasets_and_Benchmarks_Track.html.

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 2019, Sydney, Australia, September 20-25, 2019*, pp. 947–952. IEEE, 2019. doi: 10.1109/ICDAR.2019.00156. URL https://doi.org/10.1109/ICDAR.2019.00156.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige,

Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs, 2025. URL https://arxiv.org/abs/2504.13958.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2025. URL https://arxiv.org/abs/2209.13085.

Zhivar Sourati, Zheng Wang, Marianne Menglin Liu, Yazhe Hu, Mengqing Guo, Sujeeth Bharadwaj, Kyu J. Han, Tao Sheng, Sujith Ravi, Morteza Dehghani, and Dan Roth. LAD-RAG: layout-aware dynamic RAG for visually-rich document understanding. *CoRR*, abs/2510.07233, 2025. doi: 10.48550/ARXIV.2510.07233. URL https://doi.org/10.48550/arXiv.2510.07233.

Li Sun, Liu He, Shuyue Jia, Yangfan He, and Chenyu You. DocAgent: An agentic framework for multi-modal long-context document understanding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17701–17716, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.893. URL https://aclanthology.org/2025.emnlp-main.893/.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6088–6109, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.310. URL https://aclanthology.org/2025.naacl-long.310/.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. General ocr theory: Towards ocr-2.0 via a unified end-to-end model, 2024. URL https://arxiv.org/abs/2409.01704.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 7909–7928. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.401. URL https://doi.org/10.18653/v1/2025.emnlp-main.401.

Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. Molorag: Bootstrapping document understanding via multi-modal logic-aware retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 14024–14045. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.708. URL https://doi.org/10.18653/v1/2025.emnlp-main.708.

Yixin Xiang, Yunshan Ma, Xiaoyu Du, Yibing Chen, Yanxin Zhang, and Jinhui Tang. MAB-DQA: addressing query aspect importance in document question answering with multi-armed bandits. *CoRR*, abs/2604.08952, 2026. doi: 10.48550/ARXIV.2604.08952. URL https://doi.org/10.48550/arXiv.2604.08952.

Hao Yan, Yuliang Liu, Xingchen Liu, Yuyi Zhang, Minghui Liao, Jihao Wu, Wei Chen, and Xiang Bai. Docseeker: Structured visual reasoning with evidence grounding for long document understanding, 2026. URL https://arxiv.org/abs/2604.12812.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yongyue Zhang and Yaxiong Wu. Mldocrag: Multimodal long-context document retrieval augmented generation, 2026. URL https://arxiv.org/abs/2602.10271.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.547. URL https://aclanthology.org/2025.findings-acl.547/.

Yuanlei Zheng, Pei Fu, Hang Li, Ziyang Wang, Yuyi Zhang, Wenyu Ruan, Xiaojin Zhang, Zhongyu Wei, Zhenbo Luo, Jian Luan, Wei Chen, and Xiang Bai. Doc-v*:coarse-to-fine interactive visual reasoning for multi-page document vqa, 2026. URL https://arxiv.org/abs/2604.13731.

Unified Type	Raw Categories from Layout Parser ($\mathcal{C}_{\text{raw}}$)
sec	Doc Title, Paragraph Title, Figure Title
text	Text, Abstract, Content, Reference, Reference Content, Footnote, Vision Footnote, Algorithm, Number, Aside Text, Vertical Text
img	Image, Seal, Chart
tab	Table
eq	Display Formula, Formula Number, Inline Formula
discarded	Header, Footer, Header Image, Footer Image

Table 4: The mapping definition ψ from the 25 raw layout categories $\mathcal{C}_{\text{raw}}$ to the unified structural types in $\mathcal{T}_{\text{type}}$. Categories in the *discarded* set are filtered out to prevent layout noise from contaminating the document tree.

A Document Tree Construction Algorithm

Algorithm 1 gives the complete pseudocode for the five-stage tree construction pipeline outlined in §3, covering layout parsing, element typing, cross-page element merging, skeleton construction, and final tree assembly.

Algorithm 1 Document Tree Construction

Require: PDF $\mathcal{D}$ with pages $\{p_1, \ldots, p_P\}$, layout parser f_{parse}, description model f_{desc}, type mapping ψ, continuity predicate $\text{Cont}(\cdot, \cdot)$
Ensure: Document tree $\mathcal{T} = (V, E)$
$\mathcal{E} \leftarrow \emptyset$ {global element pool}
for $i = 1$ to P **do**
  $(\mathcal{L}_i, m_i) \leftarrow f_{\text{parse}}(p_i)$
  for each $(b_j, c_j, x_j) \in \mathcal{L}_i$ **do**
    $\tau_j \leftarrow \psi(c_j)$
    $d_j \leftarrow f_{\text{desc}}(x_j)$ if $\tau_j \in \{\texttt{img}, \texttt{tab}, \texttt{eq}\}$ else $\perp$
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{e_j = (b_j, \tau_j, x_j, d_j)\}$
  end for
end for
for adjacent pair (e_j, e_k) on consecutive pages **do**
  if $\text{Cont}(e_j, e_k) = 1$ **then**
    replace $\{e_j, e_k\}$ in $\mathcal{E}$ with $e_j \oplus e_k$
  end if
end for
Order heading elements $\sigma_1, \ldots, \sigma_M$ by document position
Build section backbone: attach each σ_m to nearest preceding $\sigma_{m'}$ with $d(\sigma_{m'}) < d(\sigma_m)$
Attach textual leaves (text) to nearest preceding section in reading order
Align visual leaves (img, tab, eq) to title anchors via d_j and route to enclosing section
return $\mathcal{T}$

B Element Type Mapping

Table 4 specifies the surjection $\psi : \mathcal{C}_{\text{raw}} \to \mathcal{T}_{\text{type}}$ used in Stage 2 of tree construction. The raw categories Doc Title, Paragraph Title, and Figure Title are mapped to sec. Visual categories are mapped to img, tab, or eq as specified in Table 4. Auxiliary categories that carry no document-level semantics are discarded.

C Reward Hyperparameters

Table 5 lists all coefficients of the reward scheme defined in Eq. (5)–(8). The five process-reward coefficients $r_{\text{hit}}, r_{\text{base}}, r_{\text{dec}}, r_{\text{back}}, r_{\text{fmt}}$ instantiate the per-turn rewards in Eq. (6); w_J and w_F specify the convex combination of the LLM-judge score and token-level F1 inside the terminal answer

reward R^{ans}; R^{vec} rewards each ground-truth-hit `vector_retrieval` call (a separate process-reward channel from `expand_node`); C_{cap} and $C^{\text{vec}}_{\text{cap}}$ are the trajectory-wide caps applied via Eq. (8) to the `expand_node` and `vector_retrieval` channels, respectively; and the three negative terms $P_{\text{max_iter}}, P_{\text{len}}, P_{\text{no_evid}}$ are terminal penalties for trajectories that exhaust the iteration budget without answering, exceed the response-length budget, or commit to an answer with an empty evidence pool $\mathcal{E}$.

Parameter	Value	Description
w_J	0.70	Weight of the LLM judge score in the terminal answer reward R^{ans}. The judge returns correct / partially correct / incorrect, mapped to 1.0 / 0.5 / 0.0.
w_F	0.30	Weight of the token-level F1 between predicted and gold answers in R^{ans}, complementing the judge signal with a continuous lexical match score.
r_{base}	0.01	Small base reward credited for any successfully formatted tool call, ensuring that the policy is not punished for issuing legitimate intermediate actions.
r_{hit}	0.10	Per-page reward for each gold evidence page *first* covered by an `expand_node` call. Only newly covered evidence pages are credited, so repeated coverage of the same page earns no additional retrieval reward.
r_{dec}	0.05	One-shot reward triggered when `decompose_query` returns ≥ 2 sub-queries, encouraging the agent to break complex questions into structurally tractable parts.
r_{back}	0.02	Reward for each `backtrack` call. Granted unconditionally so that the policy is never penalized for abandoning a wrong branch.
r_{fmt}	-0.05	Penalty per turn that emits malformed <thinking>/<action> blocks or invalid tool arguments.
R^{vec}	0.05	Per gold page hit obtained via `vector_retrieval`; structurally analogous to r_{hit} but on a separate, smaller budget so that dense retrieval and tree navigation are credited independently.
C_{cap}	0.30	Trajectory-wide cap on the sum of all per-turn process rewards (`expand_node`-channel). Enforced by the proportional rescaling of Eq. (8); prevents process rewards from overwhelming R^{ans}.
$C^{\text{vec}}_{\text{cap}}$	0.10	Separate trajectory-wide cap on the R^{vec} channel, applied with the same proportional rescaling so that excessive vector retrieval cannot dominate the return.
$P_{\text{max_iter}}$	-0.20	Terminal penalty when the trajectory exhausts its iteration budget without ever calling `generate_answer`.
P_{len}	-0.10	Terminal penalty when the trajectory hits the response-length budget mid-generation.
$P_{\text{no_evid}}$	-0.10	Terminal penalty for committing to a final answer with an empty evidence pool $\mathcal{E}$, discouraging the agent from short-circuiting the tree.

Table 5: Reward hyperparameters for DocTrek's dense process reward scheme. Symbols match those in Eq. (5)–(8).

D Training Configurations and Hardware Details

Data construction. The SFT corpus is sampled from `vidore_v3_pharmaceuticals`, which contains 1,908 question–document pairs. We prompt Qwen3.5-Plus to roll out 5 trajectories per pair under the agent template, yielding 9,540 raw trajectories. Applying the dual filter (correct final answer *and* retrieved-page overlap with the gold evidence) retains 3,398 trajectories, a yield of 35.6%. The RL stage is run on `vidore_v3_pharmaceuticals` and `vidore_v3_hr`. In both sampling and RL rollouts, each trajectory is capped at 60 <thinking>/<action>/<information> turns. Note that the training subsets (pharmaceutical and HR documents) and the four evaluation benchmarks (academic, technical, and table-centric documents) come from disjoint domains, so all reported results reflect cross-domain generalization.

Training hyperparameters. Our training pipeline consists of a cold-start supervised fine-tuning (SFT) phase followed by group-relative policy optimization (GRPO). For the cold-start SFT, the

model is trained on the 3,398 filtered trajectories for 3 epochs. In the subsequent RL stage, GRPO training is executed for 3 epochs with a group size of $G = 8$ rollouts per prompt, a total batch size of 32, and a learning rate of 5×10^{-7}. All training experiments and rollouts are conducted and accelerated on a cluster of 24 NVIDIA H20 GPUs over approximately 120 hours (2,880 GPU-hours).

Token-level returns and group normalization. For a rollout group of G trajectories, let L_i be the token length of trajectory i and let $r_{i,k}$ be the reward at token position k defined in Eq. (7). We use a token-level discount factor of $\gamma = 1.0$, giving the undiscounted reward-to-go

$$U_{i,t} = \sum_{k=t}^{L_i} r_{i,k}, \qquad \widetilde{U}_{i,t} = \begin{cases} U_{i,t}, & t \le L_i, \\ 0, & t > L_i. \end{cases} \tag{9}$$

At each token position $t \le \max_i L_i$, we compute the group mean and standard deviation over all G trajectories, including zeros for missing positions:

$$\begin{aligned} \mu_t &= \frac{1}{G} \sum_{i=1}^{G} \widetilde{U}_{i,t}, \quad \sigma_t = \mathrm{Std}_{i=1}^{G}(\widetilde{U}_{i,t}), \\ \hat{A}_{i,t} &= \frac{U_{i,t} - \mu_t}{\sigma_t + \varepsilon_{\mathrm{norm}}}, \qquad 1 \le t \le L_i. \end{aligned} \tag{10}$$

The numerical stabilizer is $\varepsilon_{\mathrm{norm}} = 10^{-6}$, which prevents a zero denominator when the group standard deviation vanishes.

Parameter-count scope. The 8B parameter count reported for DocTrek in Table 1 refers to the trainable policy backbone. The auxiliary models described in §5.1, used for document processing, tool execution, retrieval, and training, are not included in this count.

Offline construction and online inference. Document trees are constructed once offline at 13.8 s/page and reused across queries to the same document. For a P-page document with Q queries, this corresponds to an amortized preprocessing cost of $13.8P/Q$ seconds per query. After preprocessing, inference on MMLongBench-Doc averages 124 s/query and 33.2 tool calls, measured on a system equipped with eight NVIDIA H20 GPUs. We do not directly compare wall-clock costs with baselines whose published results use different hardware and system configurations.

E Retrieval Quality Analysis

To complement the answer-quality ablation in §5.3, Table 6 reports retrieval-side metrics on MMLongBench-Doc across the same four training variants. We measure Recall@k and Precision@k at $k \in \{3, 5, 10\}$, where the retrieval pool of a trajectory is the ordered list of leaves opened by `expand_node` (and, when invoked, returned by `vector_retrieval`). Recall@k is the fraction of gold evidence pages covered by the top-k leaves; Precision@k is the fraction of those leaves that fall on gold evidence pages. Together these metrics capture how quickly each variant locates the correct evidence, how cleanly it avoids irrelevant content, and how well it ranks evidence within the retrieved set.

Variant	Recall@k			Precision@k		
	@3	@5	@10	@3	@5	@10
Training-Free	33.84	41.94	44.36	16.77	13.11	7.06
SFT	41.03	53.44	60.33	19.76	16.69	9.86
w/o Dense Reward	40.88	52.93	59.72	19.95	16.65	9.86
DocTrek	**42.30**	**56.83**	**68.85**	**20.07**	**17.76**	**11.38**

Table 6: Retrieval quality on MMLongBench-Doc. The retrieval pool is the ordered list of leaves opened by `expand_node` (and, when invoked, returned by `vector_retrieval`). Recall@k and Precision@k jointly measure evidence coverage, retrieval purity, and ranking quality within the top-k.

F PROMPTS

We list below all prompts used in DocTrek. The *main-agent system prompt* (Prompt 1) governs the reason–act–observe loop of both the cold-start sampling agent (a closed-source LLM that produces SFT trajectories) and the trained policy at RL and inference time, so the template is identical across stages. Prompt 2-5 drive the dedicated sub-models invoked from inside individual tools—`decompose_query`, `expand_node`, `backtrack`, and `generate_answer`. The two structural tools (`list_children` and `vector_retrieval`) are purely deterministic operations over the document tree and the dense index, so they do not consume an LLM prompt. Prompt 6 specifies the LLM judge used during RL training to compute the answer-correctness component of R^{ans}.

Prompt 1: Main-Agent System Prompt (used for cold-start sampling, RL training, and inference).

```
You are a document retrieval and question-answering agent. Your task is to find
relevant evidence from a document tree structure and answer the user's question.
## Available Tools
You have 6 tools: decompose_query, list_children, expand_node, backtrack,
vector_retrieval, and generate_answer.
## Output Format
Every response MUST follow this exact structure:
<thinking>
Your internal reasoning: analyze current state, decide next action.
</thinking>
<action>
{"tool_name": "tool_name_here", "arguments": {...}}
</action>
## Rules
Select only ONE child per step; never skip levels. Do not revisit visited nodes (marked
[VISITED]). The <action> block must contain valid JSON.
```

Prompt 2: `decompose_query` Sub-Model Prompt.

```
You are a query decomposition expert. Given a question about a document, decompose it
into 1-3 focused sub-queries that are optimized for searching a hierarchical document
tree.
## Rules
1. Each sub-query should be specific and self-contained.
2. Simple questions should be rewritten as 1 search-friendly sub-query.
3. Complex questions may need 2-3 sub-queries targeting different aspects.
4. Sub-queries should use concrete terms that match document section titles or content.
## Input
- Original Question: {question}
- Document Title: {document_title}
- Document Structure: {document_summary}
## Output Format (JSON)
{"reasoning": "...", "sub_queries": ["...", "...", "..."]}
Output ONLY the JSON, no other text.
```

Prompt 3: `expand_node` Relevance Judge (multimodal variant).

```
You are a relevance judge. Given a document node's content (text + image crop) and a
search query, determine whether the node contains information that helps answer the
query.
## Search Query: {sub_query}
## Node Content: type={node_type}, title={node_title}, page={page_num},
text={content}, description={description}. The cropped image of this node is attached.
## Output Format (JSON)
{"relevant": true/false, "confidence": 0.0-1.0, "reason": "..."}
Output ONLY the JSON, no other text.
```

Prompt 4: `backtrack` Decision Prompt.

You are a search strategy advisor. The agent is performing a DFS over the document tree and needs to backtrack because the current branch is exhausted or irrelevant.

Current Search Context: current sub-query, backtrack reason, current DFS path from root, per-node details.

Instructions
Determine which ancestor on the path still has unexplored branches worth investigating for the current sub-query. Prefer the nearest useful ancestor; do not go all the way back to root unless necessary.

Output Format (JSON)
{"target_node_id": <int>, "reasoning": "..."}
Output ONLY the JSON, no other text.

Prompt 5: `generate_answer` Prompt.

You are an expert Document Question-Answering assistant. Answer the user's question strictly based on the provided document evidence (text and images).

Original Question: {question}
Collected Evidence: {evidence_text} (text and image crops)

Guidelines
1. **Strictly grounded**: use ONLY the provided evidence; no external knowledge. If insufficient, the answer must be "Not answerable".
2. **Evidence handling**: filter irrelevant noise; for images/tables extract precise data points; flag conflicts in your reasoning.
3. **Answer format** inside <answer> tags, no filler phrases.
4. **Citations**: use [Page X] in internal reasoning only, not inside the answer tag.

Response Structure
First reason step-by-step (question type, evidence filtering, data extraction). Then output the final answer enclosed in <answer></answer> tags.

Prompt 6: R^{ans} LLM-Judge Prompt (used during RL training).

System. You are a strict but fair grader for a document question-answering task. Given a question, a reference ground-truth answer, and a predicted answer, judge whether the prediction correctly addresses the question, treating the ground truth as authoritative. Return STRICT JSON only, with no extra text:
{"verdict": "correct" | "partially_correct" | "incorrect", "reason": "<one short sentence>"}

User.
Question: {question}
Ground Truth: {ground_truth}
Predicted: {prediction}

Rubric
- **correct**: the prediction fully captures the key facts of the ground truth (paraphrasing is OK).
- **partially_correct**: the prediction captures part of the key facts, or is semantically related but incomplete / contains extra incorrect information.
- **incorrect**: the prediction misses the key facts, contradicts the ground truth, or is irrelevant / empty.

Now judge and return strict JSON.