DR PIXEL: PAGE-NATIVE DEEP RESEARCH WITH VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex documents such as scientific papers contain rich information in tables, figures, and equations; however, existing deep research (DR) agents typically reduce papers to OCR-extracted text. We instead ask whether vision-language models (VLMs) can enable *page-native deep research*: directly retrieving and reasoning over rendered document page images. We introduce DR PIXEL, the first end-to-end multimodal DR agent that uses rendered page images as its sole retrieval and reasoning unit, operating over an open corpus of 1.06M pages. To evaluate this setting, we construct DRPIXELBENCH, a benchmark on scientific papers requiring reasoning over text, figures, tables, and equations. Surprisingly, page-native DR already works well without training, outperforming OCR-based baselines across state-of-the-art models like GPT-5.4, Gemini-3.5-Flash, and Qwen3.6-27B, while requiring substantially fewer tool calls.

We next turn to DR training. A key challenge is that demonstration traces are inherently noisy: some reasoning and action steps are critical to reaching the correct answer, while others are redundant or failed attempts, yet SFT treats them all equally. To address this, we introduce *counterfactual credit reweighting*, which weights individual reasoning and action steps according to their contribution to the final answer. Our training improves retrieval recall by about 20% and final accuracy by over 10% on DRPIXELBENCH, and likewise outperforms the untrained baseline on all four additional benchmarks. On out-of-distribution ViDoRe-V3, DR PIXEL comes within 1.8% of an oracle that is given the gold pages directly. Our results suggest that the conventional *OCR-then-retrieve* pipeline is not only unnecessary but suboptimal: directly retrieving and reasoning over visual document pages makes complex DR both more effective and more efficient. [1]

1 INTRODUCTION

Deep research (DR) agents answer complex information-seeking questions by iteratively searching, reading, and reasoning over large collections of evidence (OpenAI, 2025; Gemini, 2025; Li et al., 2025d;a). As DR expands beyond the open web into domains such as law, finance, and science, it increasingly operates over specialized document collections (Li et al., 2025b; Choi et al., 2026; Xie et al., 2024; Chen et al., 2025a;b). These documents are inherently multimodal: critical evidence often lies not only in text, but also in tables, figures, charts, and equations.

Yet, such documents are typically reduced to plain text: PDFs are linearized into text, and retrieval and reasoning operate entirely in text space. This introduces two key limitations. First, linearization discards or distorts visual structure, making information in figures, tables, or charts difficult or impossible to recover (Poznanski et al., 2025; Lo et al., 2023). Second, extraction and OCR errors propagate into downstream retrieval and reasoning (Zhang et al., 2025a; Wang et al., 2026).

This raises a natural question: *Can deep research instead be page-native, retrieving and reasoning directly over rendered document pages?* In such an approach, each page remains an image: a visual retriever searches over rendered pages, and a vision-language model (VLM) reasons directly over the retrieved images. Prior work has shown that such page-native retrieval can match or outperform text-based document retrieval (Faysse et al., 2025; de Souza P. Moreira et al., 2026), and has extended

[1] Our data and code is in: `https://anonymous.4open.science/r/DRPixel_Submission-3848/`.

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.

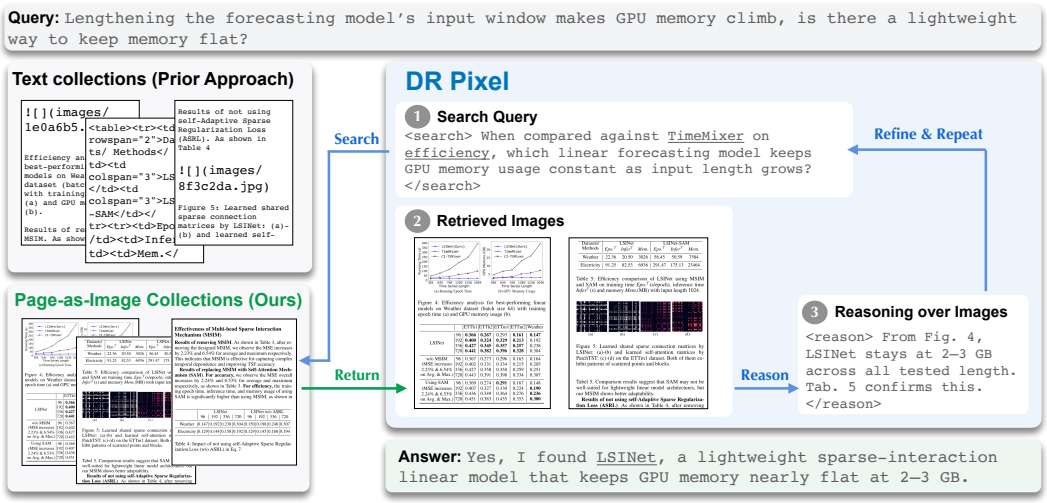

Figure 1: Overview of DR PIXEL. DR agents alternate between search, retrieval, and reasoning. Prior work operates in text space over OCR-extracted papers; DR PIXEL operates natively in visual space, with both retrieval and reasoning performed directly over rendered page images.

this approach to retrieval-augmented generation (RAG) tasks. We study this approach in the deep research regime, where answering a query requires many rounds of retrieval and reasoning. To this end, we introduce DR PIXEL, the first *page-native* multimodal DR agent that iteratively searches and reasons directly over 1.06M rendered pages from scientific papers.

The first challenge is evaluation: existing DR benchmarks are either predominantly text-based (Chen et al., 2026; Xiong et al., 2026; Asai et al., 2024) or require relatively simple retrieval where a single search already achieves high recall (Wang et al., 2025a; Tanaka et al., 2023). To bridge this gap, we introduce DRPIXELBENCH, comprising 500 questions over 1M pages of scientific papers. It combines complex deep research queries that require long chains of retrieval and reasoning with the visual reasoning demands of scientific papers, such as interpreting experimental results in figures and tables. Each question is grounded in a gold page and paired with a directly verifiable answer.

We then instantiate page-native deep research as an end-to-end VLM agent. DR PIXEL indexes rendered pages directly with a VLM embedding model (Li et al., 2026a), feeds retrieved images to the VLM agent, and maintains them as visual evidence throughout the trajectory, without ever being converted into plain text. The agent iteratively formulates queries, retrieves images, and reasons over evidence accumulated across turns to generate the final answer. Our evaluation on DRPIXELBENCH reveals a consistent advantage of page-native DR over OCR-based pipelines, even without targeted training: page-native DR improves accuracy by 10.0–19.3 points across GPT-5.4 (OpenAI, 2026), Gemini-3.5-Flash (Google, 2026), and Qwen3.6-27B (Qwen, 2026), while using 30%–50% fewer retrieval calls than OCR-based pipelines.

Still, the open-weight Qwen3.6-27B lags behind closed models (38.6 vs. 51.9), motivating us to train DR PIXEL. A key challenge is that demonstration traces for supervised fine-tuning (SFT) are often noisy, and are a mix of useful and unhelpful steps: some are critical to reaching the correct answers, while others are redundant or failed attempts. To account for this, we introduce *counterfactual credit reweighting*: a general recipe for DR training that weights each reasoning and action step by its contribution to the final answer. For each step, we estimate its contribution by removing it from the trajectory and assigning higher training weight when its removal degrades answer quality.

Our training method significantly improves the performance: with Qwen3.6-27B, it improves accuracy by 13.5 points (35% relative) over the untrained model, and closes the gap to GPT-5.4. The gains also transfer beyond DRPIXELBENCH: across three additional multimodal scientific literature benchmarks, our trained DR PIXEL remains competitive and surpasses GPT-5.4 in retrieval on two of them. Notably, on the out-of-distribution ViDoRe-V3 benchmark (Loison et al., 2026), the trained DR PIXEL comes within 1.8% of an oracle that is given the gold pages directly.

We summarize our key contributions as follows:

1. We introduce DR PIXEL, a page-native deep research agent that treats papers as collections of rendered page images and iteratively retrieves and reasons over them. Across GPT-5.4, Gemini-3.5, and Qwen3.6, DR PIXEL consistently outperforms OCR-based models that convert papers to text.
2. We introduce a new DR training method, *counterfactual credit reweighting*, which weights each reasoning and search step in the training data by its contribution to the gold answer. It improves over the untrained model on all five benchmarks, and outperforms standard SFT.
3. We release DRPIXELBENCH, consisting of complex questions that require both visual document understanding (e.g., analyzing results from tables and figures) and long chains of retrieval and reasoning, providing a challenging testbed for future work on deep research.

2 RELATED WORK

Deep research (DR) agents. DR agents tackle complex, multi-step information-seeking tasks through iterative search and synthesis (Yao et al., 2023; Li et al., 2025d;c; Shao et al., 2026). Recent work extends the pipeline to multimodal deep research by adding tools such as image search, reverse image search or cross-modality retrieval (Geng et al., 2026; Yao et al., 2026). Most of them focus on open web, where text and images are retrieved separately through two different search tools (Tao et al., 2026; Li et al., 2026b; Jiang et al., 2024). We instead target a collection of documents (e.g., scientific papers), where answering complex questions often requires jointly interpreting text, tables, and figures on the same page, making separate text and image retrieval interfaces suboptimal.

Page-native document retrieval. With rapid progress with vision-language models (VLMs), bypassing OCR pipelines and directly processing rendered document images has emerged as a promising direction (Kim et al., 2022; Lee et al., 2023; Blecher et al., 2024; Wei et al., 2026). Building on this, visual document retrieval retrieves rendered pages directly, established by ColPali and followed by increasingly capable visual retrievers (Faysse et al., 2025; Günther et al., 2025; Li et al., 2026a). More recent work extends this approach to further generate answers (Yu et al., 2025; Wang et al., 2025a; 2026) and add iterative retrieval (Wang et al., 2025b; Shen et al., 2026). However, these settings require relatively shallow search, e.g., on ViDoSeek (Wang et al., 2025a) and SlideVQA (Tanaka et al., 2023), a single retrieval step already recalls the gold page in the top five for $> 90\%$ of questions (Wang et al., 2025a; Giouroukis et al., 2025). We instead study page-native deep research where complex queries require sustained search and reasoning over 10–20 steps.

Agent training. Broadly, agents that generate long action trajectories are commonly trained with outcome-based reinforcement learning (RL) (Jin et al., 2025; Zheng et al., 2025; Wang et al., 2025b; Shen et al., 2026), sometimes with rubrics (Shao et al., 2026), or through supervised fine-tuning (SFT) on teacher-generated trajectories (Chen et al., 2023; Li et al., 2025a). Both provide limited credit assignment across a trajectory: outcome-based RL assigns a single reward to the rollout, and SFT imitates every teacher step regardless of its utility. Prior work has explored finer-grained supervision through process reward models that score each intermediate step (Lightman et al., 2023) or by leveraging token- or step-level supervision coming from a teacher model conditioned on privileged context (Meng & Chen, 2026; Wu et al., 2026), both during RL. In this paper, we propose an alternative approach that retains the simplicity of SFT: we reweight each agentic step by its counterfactual contribution to the final answer, estimated by removing the step and measuring the resulting change in answer quality.

3 DRPIXELBENCH

Existing benchmarks for complex information-seeking document collections largely fall into two categories: text-based DR benchmarks and multimodal document QA benchmarks (Table 1). Text-based deep research benchmarks evaluate iterative search and synthesis over large corpora, but often rely on templated questions (Chen et al., 2026) or long-form answers that require nondeterministic, rubric-based evaluation (Du et al., 2026; Asai et al., 2024); crucially, they are text-only. Multimodal QA benchmarks instead require visual understanding from reports or scientific papers (Ma et al., 2024; Tanaka et al., 2023; Wang et al., 2025a; Huang et al., 2026; Li et al., 2024; Shorten et al., 2026; Loison et al., 2026). However, their questions are often templated and require relatively shallow retrieval, with only one or a few steps needed to locate the relevant evidence.

We evaluate on four existing benchmarks (IRPapers, M3SciQA, AirQA, ViDoRe-V3) but argue that page-native DR requires a new benchmark, meeting four key properties: (1) *agentic search*, where relevant evidence typically requires multiple rounds of retrieval and reasoning; (2) *multimodal evidence*, where answers depend on text together with figures, tables, or equations; (3) *non-templated questions*, avoiding fixed schemas; and (4) *short, verifiable answers*, allowing direct evaluation against gold answers without relying on imperfect rubrics.

Benchmark	Agentic	MM	Non-templ.	Verif.
Deep research				
BrowseComp-Plus	✓	✗	✗	✓
DeepResearch Bench	✓	✗	✓	✗
ScholarQABench	✓	✗	✓	✗
Multimodal QA				
AirQA (comp.)	✓	✓	✗	✓
M3SciQA	✓	✓	✗	✓
IRPAPERS	✗	✓	✓	✓
ViDoSeek	✗	✓	✗	✓
SlideVQA	✗	✓	✗	✓
ViDoRe V3	✗	✓	✗	✓
DRPixelBench	✓	✓	✓	✓

Table 1: DRPIXELBENCH is the only benchmark combining iterative search over a large corpus (**Agentic**), multimodal evidence (**MM**), non-templated questions (**Non-templ.**), and verifiable short answers (**Verif.**).

DRPIXELBENCH. To this end, we build DRPIXELBENCH over scientific literature. Scientific papers are naturally multimodal, with important information distributed across text, tables, figures, and equations.

We collect papers from arXiv and the open-access proceedings of major conferences, and merge in the papers of AirQA (Huang et al., 2026), M3SciQA (Li et al., 2024) and IRPAPERS (Shorten et al., 2026) so that all four benchmarks share one corpus of 99,055 papers.

Inspired by Dasigi et al. (2021), we generate each question from a paper's abstract rather than from its full text to keep it tied to the paper's main contribution. As an abstract reports its findings with few supporting details, we first enrich it. Using Gemini-3.5-Flash (Google, 2026), we extract atomic facts from the full paper, such as a comparison in a table or a trend in a figure, and add them to the abstract. We then prompt the model to write a question from the enriched abstract alone, so that the question is driven by curiosity rather than by evidence already in view. We further ask the model to locate the supporting evidence in the full paper and produce a short-form answer. Depending on the form of this evidence, each question is labeled *Text* or *Visual*: a Text question is answered from the running text, while a Visual question requires a table, figure, or equation. Each QA pair then undergoes four automatic checks to ensure that the question is plausible, the gold evidence is sufficient, non-textual evidence is necessary, and the answer cannot be recovered from parametric knowledge alone. Finally, human experts review the question and rewrite it for naturalness. The resulting DRPIXELBENCH contains 500 QA pairs. More details, including the data sources, parsing details, filter criteria, annotation details, and data statistics, are reported in §A.

4 DR PIXEL INFERENCE

We introduce DR PIXEL: a deep research agent that uses rendered page images as both the retrieval unit and the model input. We describe our method (§4.1) and present results showing that DR PIXEL consistently outperforms OCR-based pipelines (§4.2).

4.1 INFERENCE METHOD

Following Li et al. (2025a), we formulate a DR agent as a ReAct loop (Yao et al., 2023). At each turn t, the agent produces a thought and an action, $(\tau_t, a_t) \sim \pi(\cdot \mid \mathcal{H}_{t-1})$, where $\mathcal{H}_{t-1}$ denotes the history of previous thoughts, actions, and search results. An action a_t is either a search query or a final answer. For a search action, a search tool f returns search results $f(a_t, \mathcal{D}) \subset \mathcal{D}$ where $\mathcal{D}$ is a fixed corpus that a search tool interacts with. A final answer action terminates the trajectory. Thus, the complete trajectory that generates a final answer after T tool calls is $[\mathcal{H}_T; \tau_{T+1}; a_{T+1}]$, where

$$\mathcal{H}_T = (\tau_1, a_1, f(a_1, \mathcal{D}), \ldots, \tau_T, a_T, f(a_T, \mathcal{D})). \quad (1)$$

Prior work. Prior work commonly parses documents into plain text, using document parsers or OCR, then operates the DR agent entirely in text space (Chen et al., 2026; Asai et al., 2024). Formally, $\mathcal{D}$ is a collection of parsed text documents, f is a text retrieval model (e.g., a dense retrieval

Retrieval Setup	GPT-5.4		Gemini-3.5-Flash		Qwen3.6-27B	
	Acc (↑)	# Calls (↓)	Acc (↑)	# Calls (↓)	Acc (↑)	# Calls (↓)
OCR-based pipelines						
Text	26.6	18.5	29.0	19.5	26.5	16.9
+ Figure (txt+fig)	32.6	14.5+3.8	32.4	18.4+1.0	28.6	15.2+0.8
DR PIXEL	**51.9**	**9.2**	**45.7**	**13.6**	**38.6**	**8.3**

Table 2: DR PIXEL versus OCR-based retrieval pipelines, all with no training. OCR-based pipelines retrieve OCR text chunks (*Text*), optionally with a separate search on the figure corpus (+ *Figure*). DR PIXEL retrieves and reads whole pages as images. **Acc** is answer accuracy. **# Calls** is the number of tool calls per question, split into text and figure calls in the text-and-figure setting.

model (Karpukhin et al., 2020) that indexes documents or text chunks into fixed-size vectors), and $\mathsf{f}(a_t, \mathcal{D})$ returns text chunks most relevant to query a_t. The agent π can be a text-only model.

However, a key limitation of this design is the lossy conversion from documents to text (Niu et al., 2026). Tables and equations may be parsed incorrectly; figures are often reduced to captions or omitted entirely; and OCR or parsing errors can corrupt even text content. Once discarded during parsing, this information cannot be recovered by downstream retrieval or reasoning.

Some prior work mitigates this limitation by adding a second, image-retrieval tool and allowing the agent to choose between text and image search (Suri et al., 2025). However, this is still limited because text and visual content are treated as separate spaces without preserving relationships between them, e.g., a figure and the surrounding text that explains it. Moreover, information loss or corruption during text parsing still remains.

DR PIXEL. Our DR agent, DR PIXEL, is an alternative instantiation of the DR agent in Equation 1: it preserves the ReAct loop but replaces the search tool f and its corpus $\mathcal{D}$ to be image-based, as illustrated in Figure 1. Offline, each paper is rendered page by page, so $\mathcal{D}$ becomes a collection of page images. Each image is encoded into a fixed-size vector using a VLM-based embedding model, e.g., Qwen3-VL-Embedding-8B. At inference time, f maps a text query a_t into the same embedding space and returns the top-k page images. These images are passed directly to the agent π, which is itself a VLM. The agent reasons over those images and generates the next thought τ_{t+1} and action a_{t+1}, repeating this process until it produces a final answer.

4.2 RESULTS

We now compare DR PIXEL against two OCR-based baselines—text-only and text+figure from §4.1—on DRPIXELBENCH.

Setup. We evaluate GPT-5.4 (OpenAI, 2026), Gemini-3.5-Flash (Google, 2026), and Qwen3.6-27B (Qwen, 2026) with `reasoning_effort=high` for the first two and `enable_thinking=True` for Qwen3.6-27B. All models run in the same ReAct loop with at most $T = 20$ searches, and all observations remain in context. We report accuracy judged by GPT-5.4-mini (OpenAI, 2026) under a strict prompt, taking a majority vote over three independent judgments.

From the same 99,055 papers, we build three corpora, one for each setting. The text-only baseline parses each PDF into text, preserving tables as HTML and equations as LaTeX, and retrieves $k = 5$ chunks of 256 words, with Qwen3-Embedding-8B (Zhang et al., 2025b). The text+figure baseline additionally retrieves $k = 3$ extracted figures when the image-search tool is invoked, using Qwen3-VL-Embedding-8B (Li et al., 2026a). DR PIXEL, instead, searches 1,062,153 page images rendered directly from the PDFs, returning $k = 3$ pages, also using Qwen3-VL-Embedding-8B. We choose k for each setting to approximately match the agent's retrieved token budget (Table 14 for details).

Page-native deep research outperform OCR-based pipelines. As shown in Table 2, DR PIXEL consistently outperforms both OCR-based settings across all three models, improving the text-and-figure pipeline over accuracy by 19.3, 13.3, and 10.0 points for GPT-5.4, Gemini-3.5-Flash, and Qwen3.6-27B respectively. Adding a figure-search tool yields only marginal gains, as agents rarely invoke it despite explicit instructions in the prompt. DR PIXEL is also more efficient, reducing retrieval calls by 30% to 50%. We further break down GPT-5.4's accuracy by evidence source

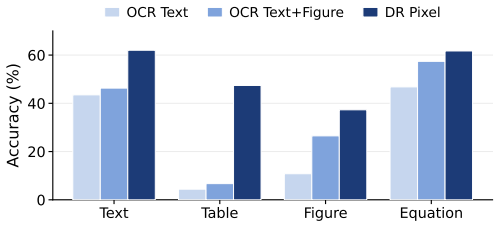

Table 3: GPT-5.4 accuracy by evidence source under the two OCR-based settings and DR PIXEL. DR PIXEL consistently outperforms OCR-based pipelines, even on Text questions.

	GPT-5.4	Qwen3.6-27B	Δ
Retrieval			
Page hit	45.7	33.8	−11.9
Repeated queries	0.0	5.3	+5.3
Reading			
Acc | page hit	71.1	63.9	−7.2
hit on last search	74.4	67.4	−7.0
hit on earlier search	69.7	60.2	−9.5

Table 4: Decomposing gap between GPT-5.4 and Qwen3.6-27B. Qwen3.6-27B retrieves the gold page less often and uses it less well once found.

in Figure 3. DR PIXEL achieves the highest accuracy across all four evidence types, including questions that require only textual evidence.

Open-weight models lag in both retrieval and reasoning. Although Qwen3.6-27B issues a similar number of retrieval calls to GPT-5.4, it trails by 13.3 points in accuracy under DR PIXEL setting. As shown in Table 4, we decompose this gap into retrieval and reasoning. In retrieval, Qwen3.6-27B retrieves the gold page 11.9% less often, and its searches are less productive with 5.3% of its queries duplicate an earlier one and 19.9% of its trajectories contain at least one duplicate. In reasoning, Qwen3.6-27B struggles to use pages retrieved earlier in the trajectory. When the gold page comes from the latest search, it trails GPT-5.4 by only 7.0 points, but the gap widens to 9.5 when the page was retrieved in an earlier search. This motivates training an open-weight model, as described in §5.

5 DR PIXEL TRAINING

Having shown that page-native DR improves performance even without training, we next train DR PIXEL to further improve its capabilities. We denote the untrained model as DR $\text{PIXEL}_{\text{Base}}$ and the trained model as DR $\text{PIXEL}_{\text{Trained}}$. We first describe the training method (§5.1), then present the main results (§5.2) and ablations (§5.3).

5.1 TRAINING METHOD

We train DR PIXEL with supervised fine-tuning (SFT). We sample trajectories from GPT-5.4 on training questions generated using the same pipeline as §3, but without human rewriting or verification; and use them as training data.

However, we find standard SFT has a problem because the trajectories in the training data often contain repeated or low-value searches that do not contribute to the final answer. Naively imitating these unhelpful steps encourages excessive searching and inefficient use of the retrieval budget. Inspired by process rewards, which assign credit to individual intermediate steps (Lightman et al., 2024; Setlur et al., 2025; Zhou et al., 2024; 2025), we instead weight each step by its contribution to the final answer. Specifically, our *counterfactual credit reweighting* removes each step from a fixed teacher trajectory and measures the resulting drop in the gold-answer likelihood under the frozen base model, assigning higher weight to steps that are more consequential.

Counterfactual credit. Formally, let $y^\star$ denote the gold answer, $\mathcal{H}_i$ the history after i search calls as defined in Equation 1, and π_0 the frozen base policy. For any history $\mathcal{H}$, we measure how strongly it supports the gold answer by the log-likelihood that π_0 assigns to $y^\star$:

$$\ell(\mathcal{H}) = \log \pi_0(y^\star \mid \mathcal{H}) \,. \tag{2}$$

We assign credit separately to reasoning and search steps. For a reasoning step τ_i, we measure how much removing it from $\mathcal{H}_T$ decreases support for the gold answer. For a search step a_i, we measure how much the search results $f(a_i, \mathcal{D})$ increase this support relative to the history before the search:

$$c_i^\tau = \ell(\mathcal{H}_T) - \ell(\mathcal{H}_T \setminus \tau_i) \,, \qquad c_i^a = \ell(\mathcal{H}_i) - \ell(\mathcal{H}_{i-1}). \tag{3}$$

Thus, c_i^τ captures the contribution of a reasoning step to organizing the accumulated evidence, while c_i^a captures the value of the new evidence returned by a search.

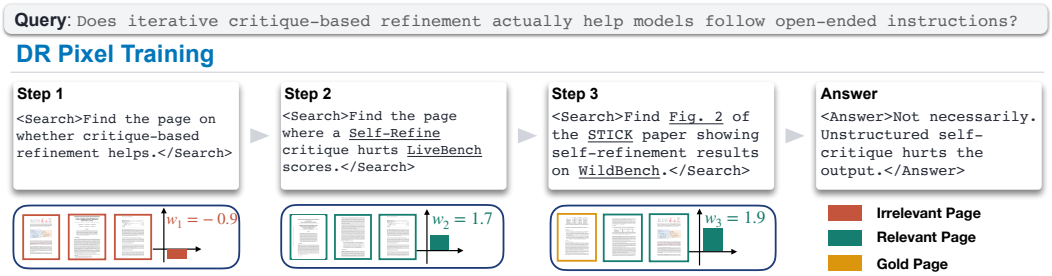

Figure 2: An illustration of *counterfactual credit weighting* (§5.1). The first search returns three off-topic pages and lowers the gold score, so the step has a negative score and weight 0. The second retrieves the gold paper, though not the exact gold page. This increases the gold score and receives positive weight. The third retrieves the exact evidence pages and likewise retrieves positive weight.

Weight assignment and normalization. Because raw step credits are unbounded, we convert them into rank-based weights for each step type $u \in \{\tau, a\}$. Specifically, steps with non-positive credit receive weight 0. For every positive credit, if c_i^u is the k-th smallest among n surviving steps of type u, we define $\text{rank}(c_i^u) = k/n$. We then multiply by 2 so the average positive weight remains near 1, matching the scale of standard SFT. The resulting weight w_i^u is:

$$w_i^u = \begin{cases} 0, & c_i^u \leqslant 0, \\ 2 \cdot \text{rank}(c_i^u), & \text{otherwise.} \end{cases} \tag{4}$$

Because each reasoning step directly produces the following search query, we further couple their weights. If a search query is masked, we also mask its preceding reasoning step; if a reasoning step has zero credit toward the final answer but increases the likelihood of the next search query over other steps, we assign it the mean positive weight $\bar{w}$. We denote the adjusted reasoning weight by $\tilde{w}_i^\tau$.

The final training objective is a token-level weighted loss, with a constant weight λ on the final-answer tokens:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{T-1} \left[\tilde{w}_i^\tau \sum_{t \in \tau_i} \log \pi_\theta(x_t \mid x_{<t}) + w_i^a \sum_{t \in a_i} \log \pi_\theta(x_t \mid x_{<t}) \right] - \lambda \sum_{t \in a_T} \log \pi_\theta(x_t \mid x_{<t}). \tag{5}$$

Training details. We fine-tune Qwen3.6-27B with LoRA on about 2,500 teacher trajectories from GPT-5.4 for one epoch, using a learning rate of 1e−4, an effective batch size of 16, and answer weight $\lambda = 1.5$, on four H200 GPUs. More details are provided in §C.

5.2 Results

Setup. We evaluate on DRPixelBench and three external multimodal scientific literature benchmarks: IRPAPERS (Shorten et al., 2026), M3SciQA (Li et al., 2024), and AirQA (Huang et al., 2026). All four share a unified corpus, as we form the corpus as a union of corpora from each benchmark (§3). To test out-of-distribution generalization, we further evaluate on ViDoRe-V3 (Loison et al., 2026) over its own document corpus, with two search tools: page retrieval as in DR Pixel, and text retrieval as in the text-only setting of §4.2, where each page is parsed into text and retrieved with Qwen3-Embedding-8B. We compare DR $\text{Pixel}_{\text{Trained}}$ with GPT-5.4, Gemini-3.5-Flash, the untrained DR $\text{Pixel}_{\text{Base}}$, and vanilla SFT, which is trained on the same teacher trajectories with every token weighted equally. All agents follow the configuration in §4.2.

Model	Tool/q	PgHit	Acc|Pg	Acc
Open-weight (Qwen3.6-27B)				
DR $\text{Pixel}_{\text{Base}}$	8.3	33.8	63.9	38.6
+ vanilla SFT	14.0	**53.6**	67.2	49.0
DR $\text{Pixel}_{\text{Trained}}$	8.9	<u>52.0</u>	67.7	**52.1**
Proprietary				
GPT-5.4	9.2	45.7	<u>71.2</u>	<u>51.9</u>
Gemini-3.5-Flash	13.6	35.4	**74.1**	45.7

Table 5: Results on DRPixelBench. **Tool/q** is searches per question. **PgHit** is the percentage of questions that retrieves the gold page, and **Acc|Pg** is accuracy restricted to those questions. Results for GPT-5.4 and DR $\text{Pixel}_{\text{Trained}}$ are averaged over three runs. DR $\text{Pixel}_{\text{Trained}}$ matches GPT-5.4 in accuracy, retrieving the gold pages more often while using fewer search tool calls.

Model	IRPapers				M3SciQA				AirQA			
	T/q	DocHit	Acc|Doc	Acc	T/q	DocHit	Acc|Doc	Acc	T/q	DocHit	Acc|Doc	Acc
Open-weight (Qwen3.6-27B)												
DR PIXEL$_{\text{Base}}$	5.6	89.4	68.8	65.4	7.8	75.3	63.4	49.8	7.9	84.3	46.0	42.7
DR PIXEL$_{\text{Trained}}$	5.0	**95.6**	**82.0**	**79.4**	9.4	**90.3**	69.6	64.2	7.0	**89.7**	55.6	52.3
Proprietary												
GPT-5.4	5.4	92.2	80.1	77.8	7.7	**90.3**	**73.3**	**67.5**	6.2	87.9	**59.5**	**56.6**
Gemini-3.5-Flash	8.4	93.3	70.1	67.6	12.4	87.2	68.3	60.8	11.1	87.5	50.0	47.0

Table 6: Results on external multimodal scientific-QA benchmarks. **T/q** is searches per question. **DocHit** is the percentage of questions for which the gold document is retrieved, and **Acc|DH** is accuracy restricted to those questions. DR PIXEL$_{\text{Trained}}$ improves over DR PIXEL$_{\text{Base}}$ on all three benchmarks and adapts search effort dynamically.

Retriever	Mod.	Generator	English					French			Global
			C.S.	Fin.	Phar.	H.R.	Ind.	Phys.	Ener.	Fin.	
Oracle*	Text	Gemini 3 Pro	80.9	70.2	71.4	72.3	66.4	71.2	69.2	62.8	70.6
	Page		86.5	70.6	76.1	71.1	68.2	74.5	69.8	64.1	72.6
Qwen3-Emb-8B	Text	DR PIXEL$_{\text{Base}}$	86.5	64.7	60.7	63.8	53.7	77.4	66.6	51.9	65.7
		DR PIXEL$_{\text{Trained}}$	88.8	**71.2**	67.3	65.4	**66.1**	77.5	69.5	52.8	69.8
Qwen3-VL-Emb-8B	Page	DR PIXEL$_{\text{Base}}$	87.9	64.4	67.9	64.2	59.7	76.2	62.7	50.9	66.7
		DR PIXEL$_{\text{Trained}}$	**89.3**	67.0	**73.1**	**67.0**	65.4	**79.1**	**70.1**	**55.3**	**70.8**

Table 7: End-to-end accuracy on ViDoRe V3. Oracle* is ViDoRe V3's published upper bound. DR PIXEL$_{\text{Trained}}$ improves over DR PIXEL$_{\text{Base}}$ with both page and text retrieval on documents outside its training domain, so the training generalizes well.

DR PIXEL$_{\text{Trained}}$ matches proprietary models. As shown in Table 5, training DR PIXEL significantly improves performance on DRPIXELBENCH, increasing accuracy from 38.6 to 52.1 and even slightly surpassing GPT-5.4. Our trajectory analysis shows that the gains are concentrated in retrieval: DR PIXEL$_{\text{Trained}}$ retrieves the gold page 6.3% more often than GPT-5.4. This suggests that post-training better aligns query generation with the retriever's embedding space, therefore producing queries that are more effective for visual retrieval.

Counterfactual credit reweighting makes search more efficient. Table 5 shows that our counterfactual credit reweighting consistently outperforms vanilla SFT, primarily by suppressing the ineffective searches from teacher trajectories. The vanilla SFT achieves a slightly higher page-hit rate than DR PIXEL$_{\text{Trained}}$, but requires 57% more searches. The key difference is search efficiency: vanilla SFT exhausts the 20-call budget on 35.8% of questions, compared with only 9.8% for DR PIXEL$_{\text{Trained}}$. This is even lower than GPT-5.4 (17.0%) and the untrained backbone (19.2%).

DR PIXEL$_{\text{Trained}}$ generalizes to external multimodal QA benchmarks. As shown in Table 6, the behavior learned on DRPIXELBENCH transfers to IRPapers, M3SciQA, and AirQA. DR PIXEL$_{\text{Trained}}$ improves over the untrained backbone on all three, with accuracy gains ranging from 9.6 to 14.4 points. It also adapts its search effort to retrieval difficulty variations: compared with the backbone, DR PIXEL$_{\text{Trained}}$ searches less on IRPapers and AirQA but more on M3SciQA, where the backbone has the largest document-retrieval gap relative to GPT-5.4 (15.0%).

DR PIXEL$_{\text{Trained}}$ further generalizes across domains and retrieval modalities. ViDoRe-V3 provides a strong out-of-distribution test, covering non-scientific documents from domains such as finance and human resources. As shown in Table 7, post-training improves both page-based and text-based retrieval by 4.1 points, despite using only page-native teacher trajectories, suggesting that the learned search behavior transfers across retrieval modalities. The two settings stay close because only 1.7% of ViDoRe-V3 questions genuinely require visual content to answer. We further compare against oracle retrieval settings, where the relevant document or page is provided directly. With page retrieval, DR PIXEL$_{\text{Trained}}$ matches the Text Oracle and remains competitive with the Page Oracle.

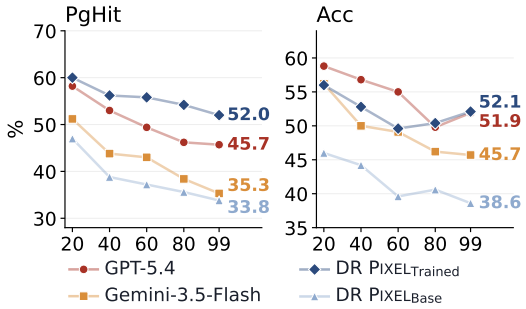

Figure 3: Effect of corpus size on retrieval and final answer accuracy, as the number of indexed documents grows from 20,000 to 99,055. DR PIXEL$_{\text{Trained}}$ degrades the most slowly in retrieval.

Setup	Tool/q	PgHit	Acc|Pg	Acc
DR PIXEL$_{\text{Base}}$	8.3	33.8	63.9	38.6
+ Text Questions	5.0	47.4	60.3	46.0
+ Visual Questions	10.1	**52.2**	**65.5**	**50.2**

Table 8: Effect of fine-tuning data type with 2,000 trajectories in each setting. *Text* and *Visual* are defined in §3. Fine-tuning on Visual questions yields stronger overall performance.

5.3 Ablation

We next ablate on a couple of key factors: corpus size, training data modalities and page resolutions.

DR PIXEL$_{\text{Trained}}$ is robust to corpus size. We scale the corpus from 20,000 to 99,000 papers while keeping the questions and gold documents fixed. As shown in Figure 3, page-hit rate declines for all models as the corpus grows, but DR PIXEL$_{\text{Trained}}$ degrades more slowly, dropping by only 8.0 points compared with 12.5 for GPT-5.4. As a result, its retrieval advantage over GPT-5.4 widens from 1.8 points at 20,000 papers to 6.3 points at 99,000, leading to final answer accuracy improvement.

Fine-tuning on *Visual* questions is more effective. The 2,500 trajectories used to fine-tune DR PIXEL$_{\text{Trained}}$ contain only *Visual* questions, as defined in §3. To justify this choice, we generate *Text* questions with the same pipeline, collect their trajectories as described in §C, and fine-tune the same base model on either 2,000 *Text* or 2,000 *Visual* trajectories, keeping all other settings fixed. As shown in Table 8, fine-tuning on Text questions causes the agent to stop earlier and retrieve the gold page less often, yielding 46.0 accuracy versus 50.2 with Visual fine-tuning. Our further analysis indicates that visual fine-tuning also transfers better: it remains within 1.4 points of Text fine-tuning on *Text* questions, while Text fine-tuning trails by 6–11 points on *Visual* questions.

Downgrading page resolution buys efficiency at a modest cost. In page-native deep research, page resolution determines the trade-off between visual fidelity and inference cost. Since processing high-resolution page images is expensive, we cap each rendered page at 0.60 megapixels, roughly 29% of its native resolution. As shown in Figure 7, this reduces input tokens by 58% and roughly halves inference cost, while accuracy drops by only 2.1 points for Gemini-3.5-Flash and by at most 9.4 points for the other models. Even at this lower resolution, page retrieval consistently outperforms text-only retrieval.

6 Conclusion

Deep research over complex documents typically begins with OCR. DR PIXEL instead performs retrieval and reasoning directly over rendered pages. We study this setting with DRPIXELBENCH, containing 500 verifiable questions over 1.06M pages of scientific literature, and show that page-native deep research consistently outperforms OCR-based pipelines across all tested models while using fewer tool calls, even without task-specific training. To close the remaining gap for open-weight models, we introduce *counterfactual credit reweighting*, which weights teacher steps by their contribution to the gold answer and brings a 27B model to GPT-5.4-level accuracy with lower search cost. These gains transfer to external benchmarks and non-scientific documents. More broadly, page-native deep research requires only document rendering rather than format-specific parsing, making it particularly attractive for private, layout-rich corpora such as financial or law documents where maintaining OCR and parsing pipelines is costly and brittle.

AI USAGE STATEMENT

Large language models are part of our method and evaluation, as described in the paper. Gemini-3.5-Flash generates and filters the questions for both the benchmark and the training data, as described in §3, and every retained benchmark question is reviewed and repaired by human experts. GPT-5.4 serves as the teacher that produces training trajectories, and GPT-5.4-mini serves as the answer judge. Beyond these uses, we used LLM-based assistants to identify relevant literature, polish the writing, check grammar, and assist with code implementation. The research ideas, experimental design, and analysis of results are the authors' own, and generative AI was not used to develop theoretical frameworks or formulate mathematical claims.

REFERENCES

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024. URL https://arxiv.org/abs/2411.14199.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fUtxNAKpdV.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023. URL https://arxiv.org/abs/2310.05915.

Shan Chen, Pedro José Ferreira Moreira, Yuxin Xiao, Samuel Schmidgall, Jeremy L. Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle Bitterman. Medbrowsecomp: Benchmarking medical deep research and computer use. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*, 2025a. URL https://openreview.net/forum?id=pSDvq2ZgPk.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-plus: A fair and disentangled evaluation benchmark for deep search agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1023. URL https://aclanthology.org/2026.acl-long.1023/.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=6z4YKr0GK6.

Chanyeol Choi, Jihoon Kwon, Alejandro Lopez-Lira, Chaewoon Kim, Minjae Kim, Juneha Hwang, Jaeseon Ha, Hojun Choi, Suyeol Yun, Yongjin Kim, and Yongjae Lee. Finagentbench: A benchmark dataset for agentic retrieval in financial question answering. In *NeurIPS 2025 Workshop: Generative AI in Finance*, 2026. URL https://openreview.net/forum?id=BIyBik7J5u.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Gabriel de Souza P. Moreira, Ronay Ak, Mengyao Xu, Oliver Holworthy, Benedikt Schifferer, Zhiding Yu, Yauhen Babakhin, Radek Osmulski, Jiarui Cai, Ryan Chesler, Bo Liu, and Even Oldridge. Nemotron colembed v2: Top-performing late interaction embedding models for visual document retrieval. *arXiv preprint arXiv:2602.03992*, 2026. URL https://arxiv.org/abs/2602.03992.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hQ0K2Hhq7H.

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

Gemini. Gemini deep research, December 2025. URL https://gemini.google/overview/deep-research/. Official blog post introducing Gemini Deep Research.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Kuan Li, Yida Zhao, Huifeng Yin, Yong Jiang, Pengjun Xie, Fei Huang, Huaxiu Yao, Yi R. Fung, and Jingren Zhou. Webwatcher: Breaking new frontiers of vision-language deep research agent. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8jsaazdAb3.

Petros Stylianos Giouroukis, Dimitris Dimitriadis, Dimitrios Papadopoulos, Zhenwen Shao, and Grigorios Tsoumakas. What's the best way to retrieve slides? a comparative study of multimodal, caption-based, and hybrid retrieval techniques. *arXiv preprint arXiv:2509.15211*, 2025. URL https://arxiv.org/abs/2509.15211.

Google. Gemini 3.5: frontier intelligence with action, May 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/. Official blog post introducing Gemini 3.5 models.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.mrl-main.36. URL https://aclanthology.org/2025.mrl-main.36/.

Tiancheng Huang, Ruisheng Cao, Yuxin Zhang, Zhangyi Kang, Zijian Wang, Chenrun Wang, Yijie LUO, Hang Zheng, Lirong Qian, Lu Chen, and Kai Yu. AirQA: A comprehensive QA dataset for AI research with instance-level evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RuUXnRyqcy.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024. URL https://arxiv.org/abs/2409.12959.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

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. *arXiv preprint arXiv:2111.15664*, 2022. URL https://arxiv.org/abs/2111.15664.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. *arXiv preprint arXiv:2210.03347*, 2023. URL https://arxiv.org/abs/2210.03347.

Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, Xixi Wu, Xuanzhong Chen, Yida Zhao, Zhen Zhang, Zhengwei Tao, Zhongwang Zhang, Zile Qiao, Chenxi Wang, Donglei Yu, Gang Fu, Haiyang Shen, Jiayin Yang, Jun Lin, Junkai Zhang, Kui Zeng, Li Yang, Hailong Yin, Maojia Song, Ming Yan, Minpeng Liao, Peng Xia, Qian Xiao, Rui Min, Ruixue Ding, Runnan Fang, Shaowei Chen, Shen Huang, Shihang Wang, Shihao Cai, Weizhou Shen, Xiaobin Wang, Xin Guan, Xinyu Geng, Yingcheng Shi, Yuning Wu, Zhuo Chen, Zijian Li, and Yong Jiang. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025a. URL https://arxiv.org/abs/2510.24701.

Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. M3SciQA: A multi-modal multi-document scientific QA benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. doi: 10.18653/v1/2024.findings-emnlp.904. URL https://aclanthology.org/2024.findings-emnlp.904/.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. LegalAgentBench: Evaluating LLM agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025b. URL https://aclanthology.org/2025.acl-long.116/.

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. *arXiv preprint arXiv:2601.04720*, 2026a. URL https://arxiv.org/abs/2601.04720.

Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, Chuanhao Li, Jiayi Tian, Chenchen Zhang, Tianhao Peng, Yancheng He, Jihao Gu, Hui Huang, Donghao Zhou, Yuanxing Zhang, Jian Yang, Ge Zhang, Wenhao Huang, Zhaoxiang Zhang, Qiangpeng Yang, and Shilei Wen. Mm-browsecomp: A comprehensive benchmark for multimodal browsing agents. *arXiv preprint arXiv:2508.13186*, 2026b. URL https://arxiv.org/abs/2508.13186.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025c. doi: 10.18653/v1/2025.emnlp-main.276. URL https://aclanthology.org/2025.emnlp-main.276/.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025d. URL https://openreview.net/forum?id=7LKKHBAMzH.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu Kang, Egor Klevak, Bailey Kuehl, Michael Langan,

Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. *arXiv preprint arXiv:2303.14334*, 2023. URL https://arxiv.org/abs/2303.14334.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel de Souza P. Moreira, Bo Liu, Manuel Faysse, Celine Hudelot, and Gautier Viaud. ViDoRe v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. doi: 10.18653/v1/2026.acl-long.755. URL https://aclanthology.org/2026.acl-long.755/.

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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=loJM1acwzf.

Zibin Meng and Kani Chen. Craft: Counterfactual credit assignment from free sibling rollouts for self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2606.29476*, 2026. URL https://arxiv.org/abs/2606.29476.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Zirui Tang, Boyu Niu, Yuanhong Zheng, Dongsheng Ma, Ziyang Miao, Hejun Dong, Siyi Qian, Junyuan Zhang, Fangdong Wang, Jingzhou Chen, Xiaomeng Zhao, Liqun Wei, Wei Li, Shasha Wang, RuiLiang Xu, Yuanyuan Cao, Lu Chen, Qianqian Wu, Huaiyu Gu, Lindong Lu, Dechen Lin, Shenguanlin, Xuanhe Zhou, Linfeng Zhang, Yuhang Zang, Xiaoyi Dong, Jiaqi Wang, Bo Zhang, LEI BAI, Pei Chu, Weijia Li, Jiang Wu, Lijun Wu, Zhenxiang Li, Guangyu Wang, Zhongying Tu, Chao Xu, Kai Chen, Bowen Zhou, Dahua Lin, Wentao Zhang, and Conghui He. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. In *The 64th Annual Meeting of the Association for Computational Linguistics – Industry Track*, 2026. URL https://openreview.net/forum?id=GqfnTTsSFr.

OpenAI. Introducing deep research, February 2025. URL https://openai.com/index/introducing-deep-research/. Official blog post introducing OpenAI Deep Research.

OpenAI. Introducing gpt-5.4, March 2026. URL https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/. Official blog post introducing GPT-5.4 models.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmOCR: Unlocking trillions of tokens in PDFs with vision language models. In *Championing Open-source DEvelopment in ML Workshop @ ICML25*, 2025. URL https://openreview.net/forum?id=0RoEHuBzc5.

Qwen. Qwen3.6-27b: Flagship-level coding in a 27b dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b. Official blog post introducing Qwen3.6 family.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=A6Y7AqlzLW.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. DR tulu: Reinforcement learning with evolving rubrics for deep research. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=97NEP1pyS3.

Yucheng Shen, Jiulong Wu, Jizhou Huang, Dawei Yin, Lingyong Yan, and Min Cao. Visor: Agentic visual retrieval-augmented generation via iterative search and over-horizon reasoning. *arXiv preprint arXiv:2604.09508*, 2026. URL https://arxiv.org/abs/2604.09508.

Connor Shorten, Augustas Skaburskas, Daniel M. Jones, Charles Pierse, Roberto Esposito, John Trengrove, Etienne Dilocker, and Bob van Luijt. Irpapers: A visual document benchmark for scientific retrieval and question answering. *arXiv preprint arXiv:2602.17687*, 2026. URL https://arxiv.org/abs/2602.17687.

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 *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)*, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.310. URL https://aclanthology.org/2025.naacl-long.310/.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. *arXiv preprint arXiv:2301.04883*, 2023. URL https://arxiv.org/abs/2301.04883.

Xijia Tao, Teng Yihua, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. MMSearch-plus: Benchmarking provenance-aware search for multimodal browsing agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VGYgG2GH0d.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. ViDoRAG: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.464. URL https://aclanthology.org/2025.emnlp-main.464/.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. VRAG-RL: Empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=EeAHhNwXPV.

Yichuan Wang, Zhifei Li, Zirui Wang, Paul Teiletche, Lesheng Jin, Matei Zaharia, Joseph E. Gonzalez, and Sewon Min. Pixelrag: Web screenshots beat text for retrieval-augmented generation. *arXiv preprint arXiv:2606.28344*, 2026. URL https://arxiv.org/abs/2606.28344.

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026. URL https://arxiv.org/abs/2601.20552.

Haoze Wu, Chuqiao Kuang, Tianyi Zhuang, and Xiaoguang Li. Beyond outcome rewards: Step-level self-distilled policy optimization for deep search agents. *arXiv preprint arXiv:2608.12764*, 2026. URL https://arxiv.org/abs/2608.12764.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, GUOJUN XIONG, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Huang Jiajia, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. Finben: An holistic financial benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=loDHZstVP6.

Lei Xiong, Kun Luo, Ziyi Xia, Wenbo Zhang, Jin-Ge Yao, Zheng Liu, Jingying Shao, Jianlyu Chen, Hongjin Qian, Xi Yang, Qian Yu, Hao Li, Chen Yue, Xiaan Du, Yuyang Wang, Yesheng Liu, Haiyu Xu, and Zhicheng Dou. Autoresearchbench: Benchmarking ai agents on complex scientific literature discovery. *arXiv preprint arXiv:2604.25256*, 2026. URL https://arxiv.org/abs/2604.25256.

Huanjin Yao, Qixiang Yin, Min Yang, Ziwang Zhao, Yibo Wang, Haotian Luo, Jingyi Zhang, and Jiaxing Huang. MM-deepresearch: A simple and effective multimodal agentic search baseline. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XZAOCytvKQ.

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*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zG459X3Xge.

Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. *arXiv preprint arXiv:2412.02592*, 2025a. URL https://arxiv.org/abs/2412.02592.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b. URL https://arxiv.org/abs/2506.05176.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.22. URL https://aclanthology.org/2025.emnlp-main.22/.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=b6rA0kAHT1.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*, 2025. URL https://arxiv.org/abs/2503.15478.

APPENDIX CONTENTS

A DRPIXELBENCH Detail **17**

A.1 OCR Parsing Detail 17

A.2 DRPIXELBENCH Statistics Breakdown 17

A.3 DRPIXELBENCH Verification Detail 17

A.4 DRPIXELBENCH Corpus Detail 18

A.5 DRPIXELBENCH Annotation 19

A.6 DRPIXELBENCH Retrieval Difficulty 19

B DR PIXEL Inference Detail **19**

B.1 Deployment and Engineering Detail 19

B.2 Comparison on Retrieval Budget 20

C More on DR $\text{PIXEL}_{\text{Trained}}$ **20**

C.1 DR $\text{PIXEL}_{\text{Trained}}$ Training Detail 20

C.2 DR $\text{PIXEL}_{\text{Trained}}$ Behavior Comparison with GPT-5.4 and DR $\text{PIXEL}_{\text{Base}}$ 21

C.3 DR $\text{PIXEL}_{\text{Trained}}$ Training Ablation 21

C.4 Other Variation for *Counterfactual Credit Reweighting* 22

C.5 *Counterfactual Credit Reweighting* Mechanism Explanation 23

D More on Experiment **24**

D.1 Supplement Material for Ablation 24

E Prompts **24**

A DRPIXELBENCH DETAIL

A.1 OCR PARSING DETAIL

All three corpora take the same PDFs as input and differ only in the retrieval unit they produce. We parse every PDF with MinerU-2.5-Pro (Niu et al., 2026) on A100 GPUs into typed blocks, from which both OCR-based corpora are built. The text corpus linearises the blocks in reading order, keeping equations as LaTeX and tables as HTML with their captions and representing a figure by its caption alone. Its unit is a 256-word chunk prefixed with the paper title. The figure corpus takes the cropped figure and chart images as its unit. The page-native corpus skips parsing entirely and renders each page at 150 dpi, so its unit is a full page image.

A.2 DRPIXELBENCH STATISTICS BREAKDOWN

Property	**Text**	**Table**	**Figure**	**Equation**	**All**
Questions					
Query Num	216	135	102	47	500
Query Len.	35.0	34.4	32.7	37.3	34.6
Answer Len.	39.8	50.7	44.4	37.5	43.5
Gold Page	1.00	1.00	1.00	1.00	1.00
Corpus					
Papers					99,055
Pages					1,062,153

Table 9: DRPIXELBENCH statistics. Questions are grouped by the modality of the evidence they are grounded in, and lengths are in tokens. Every question is grounded to exactly one page.

A.3 DRPIXELBENCH VERIFICATION DETAIL

In section 3, we present the verification we take to filter out low quality questions. Here, we present the total number of filters we apply.

Filter	Criterion	Remaining
Drafted questions		100.00%
Coined-name removal	The paper's own product or benchmark name is replaced by a generic class; questions where it cannot be removed are discarded.	99.07%
Self-contained reference	No dangling pointers such as "this method" or "the figure above".	97.17%
No internal codes	No enumeration of identifiers meaningful only inside the paper.	96.95%
Single focus	Exactly one answer is required; choosing among listed options counts as one such as "How does it do on ImageNet, and on COCO?".	86.15%
Answer shape	The gold matches its declared type: a comparison names a winner, a trend a direction, a regime a condition, and none is a disguised yes/no.	85.43%
Abstract-anchored	A solver shown only the paper's *original* abstract judges that a curious reader would plausibly ask the question, rejecting those reverse-engineered from the evidence.	**26.79**%
Phrasing repair	Questions whose wording cannot be made to read naturally are discarded.	24.79%
Evidence sufficiency	A solver given the gold evidence answers correctly.	**22.06**%
On-concern	Answering remains informative about the reader's original concern.	21.90%
Not memorised	A tool-free solver fails to produce the gold from parametric knowledge.	**14.91**%
Not prior-guessable	Generic ML priors—more data or scale wins, the proposed method wins, ablating a component hurts—do not predict the gold.	3.35%
Not web-findable	A search-and-visit agent fails after exhausting a four-call budget.	3.35%
Modality necessity	On figure- and table-grounded questions, a text-only solver forced through four retrieval rounds over the body text fails.	**2.82**%
Modality sufficiency	On text-grounded questions, the same solver succeeds, checking the modality label in the other direction.	2.33%

Table 10: Filters applied to drafted questions, in the order they run. *Remaining* is the share of drafted questions still alive after each filter. The four filters described in section 3 are in bold.

A.4 DRPIXELBENCH CORPUS DETAIL

We specify the venues of the papers that form DRPIXELBENCH corpus.

	Papers		Pages	
Venue	Count	Share	Count	Share
NeurIPS	21,448	21.7%	232,136	21.9%
ACL	12,870	13.0%	114,720	10.8%
CVPR	10,312	10.4%	96,869	9.1%
ICML	8,654	8.7%	92,765	8.7%
ICLR	8,561	8.6%	86,688	8.2%
EMNLP	7,946	8.0%	71,194	6.7%
ICCV	5,685	5.7%	53,913	5.1%
AAAI	4,835	4.9%	42,178	4.0%
NAACL	1,834	1.9%	16,317	1.5%
ECCV	1,687	1.7%	26,515	2.5%
Other venues	8,308	8.4%	122,281	11.5%
Preprint only (arXiv)	6,915	7.0%	106,577	10.0%
Total	99,055	100.0%	1,062,153	100.0%

Table 11: Venue composition of the 99,055-document, 1,062,153-page corpus. 93.0% of the documents appeared at a peer-reviewed venue.

A.5 DRPIXELBENCH ANNOTATION

Outcome	Count	Share
Accepted as generated	107	10.3%
Kept after editing	393	37.9%
Entering DRPIXELBENCH	500	48.3%
Rejected	536	51.7%
Reviewed	1,036	100.0%

Table 12: Human annotation of the DRPIXELBENCH. Of 1,036 generated questions reviewed, about one in ten survived untouched and over half were rejected.

As described in section 3, every question that passes all verification gates is reviewed manually before entering DRPIXELBENCH. We detail the annotation procedure here.

For each question, the annotator sees the question, the answer, the gold page, and the specific grounded block, such as a text span, figure, table, or equation. The full gold paper is also available. The annotator checks three things: (1) the question is well-formed and valid; (2) the answer is concise enough for reliable LLM grading; (3) the gold page is correct and sufficient to answer the question. Annotators may edit a question to fix these issues or reject it.

Only about one in ten questions passes annotation unchanged, and about half are rejected. The main reasons for rejection are that the gold answer contradicts the paper, that no single page suffices to answer the question, or that the answer is too open-ended to grade reliably. Annotation takes about 10 minutes per question.

A.6 DRPIXELBENCH RETRIEVAL DIFFICULTY

Target	R@1	R@5	R@10	R@30	R@100	MRR
Page	13.8	26.8	32.8	41.0	48.8	0.200

Table 13: Retrieval difficulty on DRPIXELBENCH, using Qwen3-VL-Embedding-8B.

B DR PIXEL INFERENCE DETAIL

B.1 DEPLOYMENT AND ENGINEERING DETAIL

Serving. GPT-5.4 is accessed through the OpenAI Responses API and Gemini-3.5-Flash through Vertex AI. Qwen3.6-27B, both the base model and our trained variants, is served with SGLang on four H200 GPUs in tensor-parallel mode at a context length of 262,144 tokens. Reasoning is set to `high` for GPT-5.4 and Gemini-3.5-Flash, and enabled for Qwen3.6-27B. Following the Qwen model card, which warns that greedy decoding in thinking mode induces endless repetition, the open model is sampled at temperature 0.6, top-p $= 0.95$, top-k $= 20$ with a presence penalty of 1.0; the proprietary models use provider defaults. All three run under an identical prompt, tool set and search budget.

Halting at the tool call. The agent must stop generating the moment it emits a tool call, so that the real retrieval result re-enters the context. For Qwen3.6-27B and Gemini-3.5-Flash we pass `</call_tool>` and `</answer>` as stop sequences. For GPT-5.4 the tools are declared as native functions and the API halts at the `function_call` item, with retrieved pages returned as image content in `function_call_output`.

Carrying reasoning across turns. Qwen3.6-27B writes its `<think>` block in-band, so it is part of the conversation. For the proprietary models we echo the provider's own representation verbatim, the `thought_signature` for Gemini-3.5-Flash and the encrypted reasoning item for GPT-5.4, carried in the request body with server-side retention disabled.

Context growth and prefix reuse. Every observation stays in context, so the agent can re-read a page it retrieved ten steps earlier, and the append-only prefix is matched by the server's radix cache at every turn, giving a 96% multimodal cache hit rate.

B.2 Comparison on Retrieval Budget

In §4.2, each text search returns $k = 5$ chunks of 256 words, while each DR Pixel search returns $k = 3$ pages. To check that this does not give DR Pixel more information, we run OCR on the retrieved pages and compare their text length with that of the retrieved chunks. On average, the text in three retrieved pages is 1.08× that of five retrieved chunks (Table 14), so the two settings receive a comparable text budget per search, and DR Pixel's advantage cannot be attributed to seeing more text from pages.

Setting	Retrieved unit	k	Words per search
Text	256-word chunk	5	1,257
Page-native	page image	3	1,359

Table 14: Text returned per search. Word counts are measured over 800 randomly sampled corpus documents including 22,149 chunks of text, 12,296 pages, with page text obtained by OCR. A chunk averages 251 words and a page 453.

C More on DR Pixel$_{\text{Trained}}$

C.1 DR Pixel$_{\text{Trained}}$ Training Detail

Optimization. We fine-tune Qwen3.6-27B with LoRA of rank 32 and scaling $\alpha = 64$, dropout 0.05, on every linear projection of the language model; the vision encoder stays frozen. Training runs for one epoch at a constant learning rate of 1e−4 with an effective batch size of 16, from four H200 GPUs at one sequence per device and four gradient-accumulation steps. We use `bfloat16`, FlashAttention, Liger kernels and padding-free packing, and set the maximum sequence length to 131,072 tokens.

Applying the weights. Counterfactual credits are computed offline against the frozen base policy at two forward passes per step and written into the training file as a per-message weight. Each assistant message is split at the boundary between its reasoning and its action so the two channels can carry different weights, and the trainer multiplies each message's token loss by its weight. The training step is therefore ordinary supervised fine-tuning at no extra cost, and the whole method sits in data preparation. Observation tokens are excluded from the loss in every arm.

Training data. Training questions come from the same pipeline as the benchmark, taken before human annotation, over a disjoint pool of papers, and no paper supporting a training question supports an evaluation question. From 1,164 *visual* seed questions whose gold evidence anchored on a table, figure, or equation, we sample six GPT-5.4 rollouts each under DR Pixel, discard those with an incorrect final answer, and keep at most the three shortest trajectories per question, leaving 2,500 trajectories. The teacher's trajectories are short, 5.1 searches per question, but heavy-tailed: under uniform token weighting the length seen by the gradient is $\mathbb{E}[n^2]/\mathbb{E}[n] = 9.0$ rather than 5.1, and trajectories with at least ten searches are 15% of the data but 42% of the steps. Filtering for correctness also means every retained trajectory ends with the answer in its context, which evaluation does not guarantee.

C.2 DR PIXEL$_{\text{TRAINED}}$ BEHAVIOR COMPARISON WITH GPT-5.4 AND DR PIXEL$_{\text{BASE}}$

	GPT-5.4	DR PIXEL$_{\text{Base}}$		DR PIXEL$_{\text{Trained}}$	
	Score	Score	Δ	Score	Δ
Retrieval					
Page hit	45.7	33.8	−11.9	52.0	+6.3
Repeated queries	0.0	5.3	+5.3	0.1	+0.1
Reading					
Acc | page hit	71.1	63.9	−7.2	67.7	−3.4
from last search	74.4	67.4	−7.0	87.8	+13.4
from earlier search	69.7	60.2	−9.5	66.0	−3.7

Table 15: Decomposing the gap to GPT-5.4 on DRPIXELBENCH, with Δ measured against GPT-5.4 in both blocks. **Page hit** is the share of questions whose gold page is retrieved, and **repeated queries** the share of search calls that re-issue a query already sent in the same trajectory. **Acc | page hit** is accuracy given the gold page was retrieved, split by whether the page arrived from the last search or had to survive further searches. Training reverses the retrieval gap, and the reading deficit that remains is confined to pages carried from an earlier search.

C.3 DR PIXEL$_{\text{TRAINED}}$ TRAINING ABLATION

Model	Tool/q	PgHit	Acc|Pg	Acc
Open-weight (Qwen3.6-27B)				
DR PIXEL$_{\text{Base}}$	8.3	33.8	63.9	38.6
+ vanilla SFT	14.0	**53.6**	67.2	49.0
DR PIXEL$_{\text{Trained}}$	8.9	<u>52.0</u>	67.7	**52.1**
Ablation of the credit				
+ answer weighting λ	11.6	54.9	67.2	48.9
+ counterfactual credit	11.0	51.9	68.7	50.9

Table 16: Results on DRPIXELBENCH. **Tool/q** is searches per question. **PgHit** is the percentage of questions for which the gold page is retrieved, and **Acc|Pg** is accuracy restricted to those questions. The lower block adds one component of the credit at a time on top of vanilla SFT baseline, and DR PIXEL$_{\text{Trained}}$ is the full method, which additionally corrects the reasoning and searching channels against each other. Each component removes searches while gradually benefit the final accuracy.

DR PIXEL$_{\text{Trained}}$ adds three ingredients to vanilla SFT: a constant weight $\lambda = 1.5$ on the final answer, the per-step counterfactual credit of Equation 3 converted to weights by rank, and the cross-channel correction described after Equation 4. Table 16 adds them one at a time, with every variation trained on the same 2,500 trajectories under the same recipe, so weighting is the only variable. Each ingredient removes searches, from 14.0 per question under uniform weighting to 8.9 under the full method and gradually benefit the final accuracy.

C.4 OTHER VARIATION FOR *Counterfactual Credit Reweighting*

Model	Tool/q	PgHit	Acc|Pg	Acc
Reference				
Global rank	11.0	51.9	68.7	50.9
Weight magnitude				
Mean-weight matching	10.1	50.4	63.1	46.8
Rank support				
Within-trajectory rank	10.0	50.3	70.9	50.1
Within-question rank	9.5	50.2	66.9	48.6
Trajectory contribution				
Per-trajectory mass	8.9	51.6	64.0	45.6

Table 17: Normalising the credit, on DRPIXELBENCH. Each variant changes one normalisation choice while the credits, the mask and the training recipe stay fixed. Columns are as in Table 5. Rescaling the weights costs reading accuracy, whether the mean weight is matched to the uniform baseline or the mass each trajectory contributes is equalised, while changing the pool a step is ranked against leaves the result unchanged.

Equation 4 fixes three properties that the counterfactual credit does not itself determine: the magnitude of the surviving weights, the pool of steps a given step is ranked against, and how much mass a single trajectory may contribute. Here we ablate on those three properties in Table 17 with the training recipe held fixed.

Weight magnitude. Masking every step of non-positive credit zeroes about a quarter of the steps and leaves the mean weight roughly a third below the uniform baseline, so the weighted run effectively takes smaller optimisation steps, and a constant factor on all weights is equivalent to a change of learning rate. Restoring the mean to the baseline settles this: accuracy given the gold page falls from 68.7 to 63.1, the largest reading loss in the family, while the number of searches barely moves.

Rank support. A counterfactual credit is larger on hard questions, where deleting any step moves the likelihood of the gold answer substantially, so ranking against the whole training set imports question difficulty into a step-level signal. The effect is measurable: the mean surviving weight of a trajectory ranges from 0.65 at the tenth percentile to 1.70 at the ninetieth, so a step's weight depends substantially on which trajectory it belongs to. We also try ranking a step only against the other steps of its own trajectory, and only against the steps of the same question's rollouts. Both remove the leak, because a step is then never compared with a step from a different question. Neither changes the result, so the bias is real but harmless.

Trajectory contribution. Equal mean weights do not equalise what a trajectory contributes: its total weighted mass spans a factor of about six between the tenth and ninetieth percentiles, since a long or lightly masked trajectory supplies far more weighted tokens. Equalising that mass is the most aggressive variant we tried and the weakest result, five points below the reference. This is because a light trajectory gets multiplied by a large ratio.

Numerical calibration is thus not where the credit acts. Rescaling the weights either does nothing or degrades reading, and changing the support of the rank does nothing at all. What remains is which steps the credit masks and how that mask is distributed along a trajectory.

C.5 *Counterfactual Credit Reweighting* MECHANISM EXPLANATION

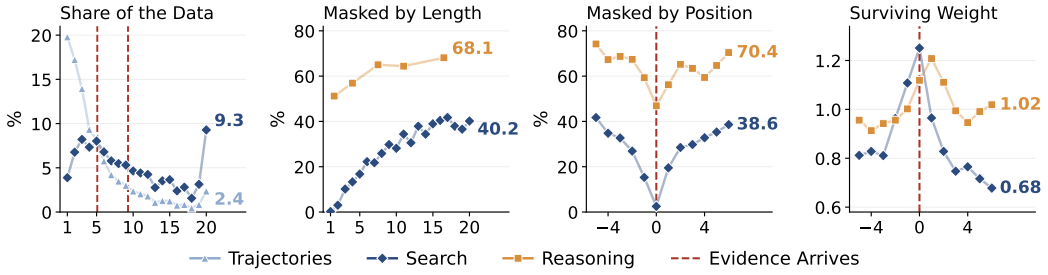

Figure 4: What the credit masks in the 2,500 training trajectories. **Share of the Data**: trajectories and search steps by trajectory length; The dashed rules mark the average trajectory, which has 5.1 searches, and the same average weighted by how many steps each trajectory contributes, which is 9.3. **Masked by Length**: mask rate against trajectory length. **Masked by Position** and **Surviving Weight**: the same two quantities against position relative to the search that first retrieves a page the final answer cites, marked by the dashed rule; that search is masked 2.5% of the time, the minimum of the panel.

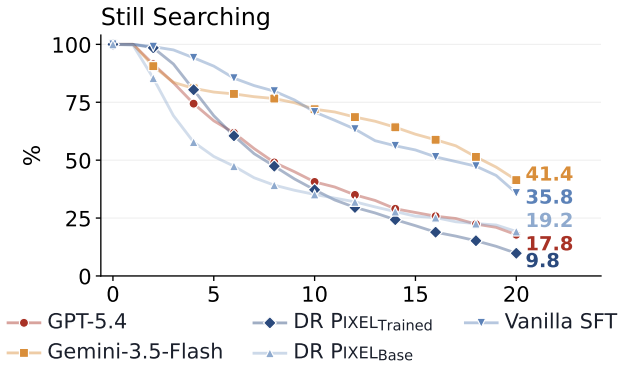

Figure 5: Share of questions that issue at least k searches on DRPIXELBENCH. The harness stops a trajectory at 20 calls, so the right-hand endpoint of each curve is that model's exhaustion rate. DR PIXEL$_{\text{Trained}}$ is the point-wise lowest curve beyond the fifth search and the only model that ends below 10%. Vanilla SFT and Gemini-3.5-Flash share a shape: both stay above 75% for the first eight searches and then run out of budget rather than deciding to stop.

Vanilla SFT and DR PIXEL$_{\text{Trained}}$ see the same trajectories and differ only in how the loss is weighted, yet one issues 14.0 searches per question and the other 8.9 at the same page hit rate. We ask what the two models do differently at inference, and what in the loss produces it.

At inference, the difference is stopping. Figure 5 plots the share of questions still searching after k calls. The curves separate at once and never cross: by the fifth search DR PIXEL$_{\text{Trained}}$ has finished a third of its questions and vanilla SFT almost none. Per step, averaged over the first twelve searches, DR PIXEL$_{\text{Trained}}$ stops with probability 10.2%, close to the untrained backbone's 9.4% and GPT-5.4's 8.9%, while vanilla SFT stops with probability 4.4%. At that rate the expected trajectory is 22.8 searches, longer than the budget allows, so the cap is what ends a third of its trajectories.

The credit is not symmetric across the two channels. The credit is much harsher on reasoning than on search. As shown in Figure 4, it masks 62.6% of reasoning steps but only 24.5% of search steps. Only search reacts to trajectory length: almost nothing is masked on trajectories with a single search, 0.2%, against 40.2% on trajectories that reach the budget, while reasoning stays between 51% and 68% whatever the length. The small slope reasoning does have is borrowed, since the cross-channel correction of §5.1 masks a reasoning step whenever its query is masked. Length is

therefore handled by the search channel: longer trajectories lose a larger share of their search steps, which is how the credit offsets the length bias, and reasoning follows along.

The credit spares the search that finds the evidence and prunes the ones after it. In each trajectory we locate the first search that returns a page the final answer cites, and measure the mask rate at each position relative to it. That search is almost never masked, 2.5% of the time, while every other search is masked between 15% and 42%. Among the steps that survive it also carries the largest weight, 1.25, against 0.7 to 0.8 for searches a few steps later. Reasoning peaks one step further on, at 1.21. This position is never supplied to the method, so the pattern is emergent. Because almost every step the credit removes is issued after the evidence is already in context, the pruning suppresses continuation rather than retrieval, which is why the search count falls while the page hit rate does not.

D More on Experiment

D.1 Supplement Material for Ablation

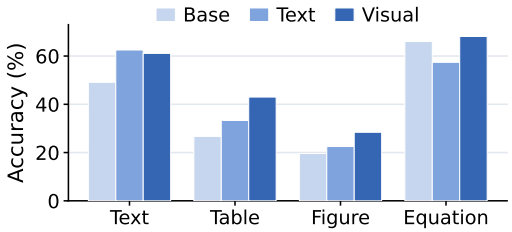

Figure 6: Each training mixture improves the question type it matches, with *Text* questions helping text questions and *Visual* questions helping multimodal ones.

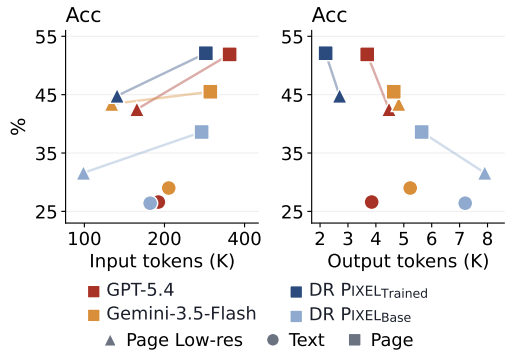

Figure 7: Accuracy against input tokens as page resolution varies. Lowering resolution reduces token cost substantially while accuracy falls only modestly, and page retrieval remains ahead of the OCR-based settings.

E Prompts

Agent system prompt

Image input

```
{retrieved page images, appended to every tool result}
```

```
You are a research assistant who answers questions about scientific papers
through iterative reasoning and research.

## Process
- Show your reasoning inside <think></think> tags.
- Call a tool with exactly:  <call_tool name="TOOL_NAME">your query</call_tool>
- When you have enough evidence, give the final answer with exactly:
  <answer>your answer</answer>
- You may alternate between thinking and tool calls several times.

## STOP after the tag  (this is the most important rule)
Each turn you emit AT MOST ONE <call_tool> or ONE <answer>, and then you STOP WRITING.
Do NOT write a <tool_output> block yourself -- WE run the tool and send you the real
results. Anything you write after the closing </call_tool> is discarded, and any tool
output you invent is a hallucination: the documents, page numbers and numbers in it
are not real.

Your turn looks like exactly this and nothing more:

    <think>What I know so far, and what I still need.</think>
    <call_tool name="TOOL_NAME">my query</call_tool>

Your <think> stays visible to you for the rest of the session, so record in it any
concrete value you read off an image -- that is your working memory.

## Tools (use ONLY these)
{tools}

## Tool Output
After each tool call the results come back to you in a <tool_output> block:
{obs}
Results from EVERY tool call stay visible to you for the rest of the session -- you
can re-read an image you retrieved earlier.

## Before you answer -- check yourself
You may only answer when you can point to the SPECIFIC page/passage id you read the
answer off (e.g. "P2_1"). Ask yourself:
  - Is the exact value/claim the question asks for actually VISIBLE in one of the
    results I received?
  - If NO -> do NOT answer. Search again with a different query.
  - "The retrieved page is about the right topic" is NOT enough. "I remember this
    paper" is NOT enough.
Answering from memory when the evidence is not in front of you is the single worst
failure mode here, because the answer will look confident and be about the wrong paper.

## CRITICAL ANSWERING RULES
- Answer ONLY from evidence you actually retrieved this session. Do NOT answer from
  prior knowledge.
- When the answer involves a specific number, score, magnitude, threshold, or a named
  model/method/dataset, you MUST retrieve it and state the SPECIFIC value(s). A vague
  or directional answer ("generally better", "tends to improve", "competitive",
  "around X") when a precise value exists is treated as WRONG.
- NEVER guess or fabricate numbers, model names, method names, arxiv ids or document
  ids. If after searching you still cannot find the specific value the question asks
  for, say so explicitly -- do NOT invent one.
- You start this session having retrieved NOTHING. The ONLY evidence that exists is
  what comes back inside a real <tool_output> block that WE send you. If you have not
  yet received a <tool_output>, you have no evidence and you MUST NOT answer -- search
  first.
- Recalling the paper from memory is not evidence. If your answer would be the same
  without looking at the retrieved pages, you have not done the task.
- For yes/no, comparison, trend, or regime questions: state the verdict/direction AND
  the supporting numbers behind it.

Illustrative GOOD vs BAD answers (fabricated examples, NOT the real question):
- Q "Does pruning hurt accuracy on the small model?"
  BAD : "Yes, pruning generally lowers accuracy somewhat."       (vague -> WRONG)
  GOOD: "Yes -- accuracy drops from 79.3% to 74.1% at 50% sparsity (Table 3), a
         5.2-point fall."
- Q "Which retriever wins on the multilingual benchmark, and by how much?"
  BAD : "A single general retriever is competitive across languages."
                                                          (names nothing -> WRONG)
  GOOD: "Retriever-Z leads, +1.8 nDCG@10 over the per-language baseline (Table 4)."

Rules: you may call tools at most {budget} times in total; after that you MUST answer.
Issue focused queries. Read the returned passages/images carefully.
```

Figure 8: The agent system prompt, identical for every model we evaluate. `{tools}`, `{obs}` and `{budget}` are filled at run time, with budget $= 20$. Models whose tool call is a native API item rather than a tag receive the same text with only the PROCESS block replaced.

Answer grading prompt

```
You STRICTLY grade an answer to a question about a research paper's figure/table/finding.
The GOLD states a SPECIFIC finding one can only know by reading the figure/table -- often
exact numbers, named models/methods/datasets, or a precise quantitative relationship.

GRADING RULE:
1. If the GOLD contains specific numbers, named entities (models/methods/datasets), or a
   precise magnitude, mark correct=true ONLY IF the PREDICTED answer states that SAME
   specific finding: the same numbers (within rounding) AND/OR the same named entities
   AND/OR the same precise relationship. Otherwise correct=FALSE. In particular:
   - Generic / directional answer that gets the trend right but OMITS the gold's specific
     numbers => WRONG.
   - Answer citing DIFFERENT numbers or DIFFERENT models/methods than the gold => WRONG
     (it guessed from prior knowledge, did not read the source).
   - Vague hedging ("levels off", "tends to", "generally", "competitive") in place of the
     gold's specific finding => WRONG.
2. If the GOLD has NO specific numbers/entities (a pure yes/no verdict or a purely
   qualitative direction), then a correctly matching verdict/direction is enough.
Paraphrase of the SAME specific finding is fine; extra correct detail is fine; a
contradicted or reversed direction is always WRONG.

answer_type = {atype} (verdict / comparison / trend / regime).

CALIBRATION EXAMPLES:
[CORRECT] GOLD: "No, it stays stable..." (no numbers) | PRED: "No, ... remains bounded,
  avoids collapse" -> verdict matches, gold has no specifics.
[CORRECT] GOLD: "operational factors; limiters = latent size and number of modules" |
  PRED: "operational overhead; size and number of processed tensors" -> states the same
  specific finding.
[WRONG] GOLD: "Jina-ColBERT-v2 within ~0.4 nDCG@10 of mDPR-FT on MIRACL" | PRED: "a single
  model such as CORA or mAggretriever is competitive" -> different models, none of the
  gold's specifics.
[WRONG] GOLD: "EMOVA 72B trails on RealWorldQA 71.0 vs 75.4" | PRED: "InternVL3-78B,
  Qwen-VL2.5-72B trail GPT-4o" -> different models, missing 71.0/75.4.
[WRONG] GOLD: "edge vanishes: 66.6 vs 62.7 (Qwen2.5) but 70.3 vs 70.9 (Qwen3)" | PRED:
  "the advantage diminishes" -> correct direction but NONE of the gold's specific numbers.
[WRONG] GOLD: "rises 2->4, peaks ~4, dips at 5" | PRED: "levels off / plateaus" -> vague,
  misses the specific peak/dip.

QUESTION: {q}
GOLD ANSWER: {gold}
PREDICTED ANSWER: {pred}
First state the gold's KEY specific finding, then decide.
Output ONLY: {"gold_key_finding": "<the specific finding the gold makes>",
              "correct": true or false,
              "reason": "<one sentence: does PRED state it?>"}
```

Figure 9: The grading prompt. A prediction is scored by majority of three independent votes from GPT-5.4-mini; an empty prediction is marked incorrect without being sent to the judge. Retrieval metrics do not pass through this prompt: page hit and document hit are decided by exact identifier match against the gold annotation.

Atomic fact extraction prompt (figure modality)

Image input

```
{cropped figure image}
```

```
You are extracting fine-grained factual content from a scientific paper FIGURE (the image).
You are given the figure image, its caption, and the surrounding body text.

Return STRICT JSON:
{
 "overview": "<=2 sentences on what the figure shows",
 "atomic_facts": [
   {"fact":"<one concrete, checkable fact READ FROM THE IMAGE>",
    "fact_type":"trend|value|comparison|ranking|structure|annotation",
    "stated_in_caption": true|false,
    "stated_in_nearby_text": true|false,
    "salience": 1-5}
 ]
}
Rules:
- Facts must be things visible IN THE IMAGE PIXELS (values, trends, orderings, crossover
  points, error bars, sub-figure differences, color/legend mappings).
- For each fact, judge honestly whether the SAME information is already explicitly stated
  in the caption / nearby text provided.
- Prefer facts that are NOT stated in caption/text. 4-10 facts.

CAPTION: {caption}

NEARBY TEXT: {nearby}
```

Figure 10: Atomic fact extraction, shown for the figure modality. Tables receive the same prompt with the OCR'd cell text in place of the image; equations receive the LaTeX source with no image; and body text is processed in chunks without caption or salience fields. The `stated_in_caption` and `stated_in_nearby_text` flags are what later let us keep only the facts a reader could not obtain from the surrounding prose.

Question generation prompt

```
You read this (augmented) abstract. As a working researcher, pick ONE thing you'd genuinely
want to check whose precise answer needs a figure/table/equation (a gap / head-to-head /
where-it-breaks / dependency / feasibility / exception) -- NOT a vague survey question, NOT
a trivial lookup.
BE IMAGINATIVE -- go for a NON-OBVIOUS, slightly surprising angle a curious reader of the
ABSTRACT would wonder about: an unexpected regime where the claim might break, a
counterintuitive comparison, a "what's the catch", a hidden tradeoff. Avoid the rote "does
the proposed method beat the baseline?". The thing you ask about must be something the
ABSTRACT itself raises (a claim/effect/tradeoff) -- NOT an internal experimental knob the
abstract never mentions. Express it as an expectation ("I'd expect X because Y"), pick the
answer_type, give 2-3 grep keywords appearing in the paper's figure/table facts
(dataset/method/metric words).{modhint}

AUGMENTED ABSTRACT:
{aug}

Return JSON {"folk_belief":"...","intent":"...",
             "answer_type":"comparison|trend|regime|verdict",
             "keywords":["..","..",".."]}
```

Figure 11: Question generation. The model sees only the abstract, enriched beforehand with the atomic facts of Figure 10, and never the body of the paper, so the question is driven by what a reader would wonder rather than by evidence already in view. Two further prompts, omitted here, turn the resulting expectation into a precise question with a short gold answer and then rewrite it in a researcher's voice.