# ICA: Information-Aware Credit Assignment for Visually Grounded Long-Horizon Information-Seeking Agents

Anonymous ACL submission

## Abstract

Long-horizon reinforcement learning for information seeking agents remains difficult because terminal rewards reveal whether the final answer is correct, but not which acquired information enabled it. This difficulty is amplified by text-derived webpage observations, where parsing, truncation, and summarization often produce incomplete and unstable content representations across trajectories. We propose an evidence-centric framework for web agent learning that represents information acquired through tools as identifiable units for comparison across trajectories. In particular, fetched webpages are represented as rendered snapshots, preserving layout and multimodal content as stable content-level observations. Building on these units, we introduce Information-Aware Credit Assignment, a post hoc reward propagation method that estimates turn-level utility scores from rollout success rates and assigns dense rewards to intermediate steps that introduced high-utility information. Integrated with GSPO, our method consistently improves performance on BrowseComp, GAIA, Xbench-DS, and Seal-0. Code and datasets will be released at https://anonymous.4open.science/r/ICA_MM_deepsearch-30DA.

## 1 Introduction

Recent systems built on large language models have advanced open-web information seeking toward increasingly sophisticated interactions with external web environments (He et al., 2026; Tao et al., 2025; Wu et al., 2026, 2025; Zhang et al., 2026), demonstrating strong performance on complex multi-hop question answering benchmarks (Kwiatkowski et al., 2019; Press et al., 2023; Mialon et al., 2024; Wei et al., 2025a; Chen et al., 2025; Pham et al., 2026). However, scaling such agents to long-horizon search remains difficult. In information seeking tasks, intermediate tool actions are useful mainly because they acquire external information that may later support the final answer. Terminal rewards, however, only indicate whether the final answer is correct, leaving unclear which search or fetch decisions actually contributed useful information. We therefore view long-horizon web agent learning as an evidence attribution problem, where effective optimization requires not only rewarding successful final answers, but also assigning credit to useful retrieval actions along the trajectory. To address this issue, we introduce Information-Aware Credit Assignment (ICA), which assigns credit according to the utility of the external information acquired by each tool action. Rather than treating all intermediate turns as uniformly responsible for the final outcome, ICA estimates which evidence units are associated with successful rollouts and propagates reward to the search and fetch turns that introduced them.

A key prerequisite for ICA is that acquired evidence must be identifiable and comparable across trajectories. However, existing web agents commonly rely on text-based parsing pipelines that convert webpages into linearized text sequences (He et al., 2026; Team et al., 2026c; Wu et al., 2026, 2025). Although convenient for language models, such representations often discard layout structure, visual cues, tables, and other presentation dependent information. They also make evidence identity unstable, since the same webpage may be parsed, truncated, chunked, or summarized differently across tool pipelines and context budgets. We therefore represent fetched webpages as rendered snapshots, which preserve visible content, layout, and multimodal cues while providing more consistent content-level observations for the same fetched source. This allows acquired information to be identified and compared across rollout trajectories, making visual webpage represen-

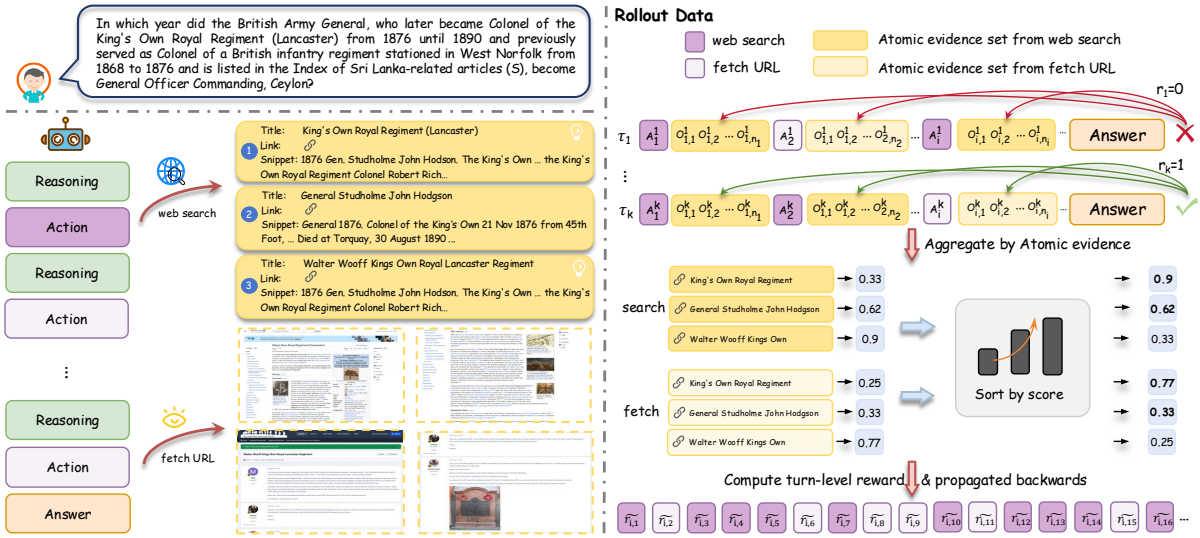



Figure 1: **ICA Framework.** On the left, an information-seeking agent alternates Reasoning-Action turns, invoking web search and fetch URL to acquire external evidence before producing a final answer. Given a batch of rollout trajectories $\{\tau_i\}$ with sparse outcome supervision, we group interactions by the webpage snapshots obtained from each visited website. We then estimate the marginal utility of each acquired webpage content by its association with successful outcomes, and propagate this signal back to the turns that revealed the content, yielding dense turn-level rewards $r_{i,t}$ for long-horizon credit assignment.

tation a basis for downstream credit assignment rather than merely an input modality choice.

Finally, as illustrated in Figure 1, we incorporate snapshots and ICA into a GSPO-based (Zheng et al., 2025) optimization pipeline and evaluate the resulting agents on a diverse suite of information-seeking benchmarks. Empirically, our approach yields consistent gains over text-based and outcome-only baselines across different model configurations, underscoring the value of grounding reinforcement learning in visually structured external observations and performing credit assignment at the granularity of acquired evidence.

In summary, this work makes the following contributions:

- We formulate long-horizon web agent learning as an evidence attribution problem, where retrieval actions are optimized through the information they introduce rather than only through terminal answer outcomes.
- We introduce rendered webpage snapshots as stable content-level observations, enabling more consistent identification and comparison of fetched information across trajectories.
- We propose Information-Aware Credit Assignment, a post hoc credit assignment method that estimates association-based utility scores for tool-acquired information and propagates dense rewards to the search and fetch turns that introduced it.
- We integrate ICA with GSPO and show consistent improvements on BrowseComp, GAIA, Xbench-DS, and Seal-0, with our 8B model outperforming most reported open-source agents under 15B parameters.

## 2 Related Work

### 2.1 Information Seeking Agents

Information seeking agents have advanced rapidly with large language models and external tool use. Proprietary systems such as DeepResearch (OpenAI, 2025a) and Grok 3 (xAI, 2025) have shown strong capabilities in open web search and complex question answering, although their training data, tool interfaces, and optimization procedures remain largely undisclosed. In parallel, open-source agents such as WebDancer (Wu et al., 2026), WebThinker (Li et al., 2025c), R1 Searcher (Song et al., 2025), and related systems adopt ReAct style interaction patterns and provide reproducible platforms for studying multi-turn search, evidence acquisition, and answer generation.

A central challenge in these systems is how external web information is represented and main-

tained during search. Some agents interact with the web through browsing operations such as search, click, and open, where clicked or opened pages are returned as textual webpage content (Lu et al., 2025). Such text observations often contain boilerplate, navigation elements, and formatting noise, which consume large context budgets. Search and summary pipelines compress webpages before reasoning (Team et al., 2026c), but summarization may remove details that later become important and introduce model-dependent variation across trajectories. Retrieval-based pipelines such as Search R1 (Jin et al., 2025) improve efficiency by chunking webpages and selecting relevant segments, yet their performance depends on chunk boundaries and reranking quality. These approaches make web information easier to consume, but they usually treat retrieved content as transient text inputs rather than identifiable information units that can be compared across trajectories. In contrast, our work represents fetched webpages as rendered snapshots and uses tool-acquired information units as the basis for assigning credit to retrieval decisions.

### 2.2 Reinforcement Learning in IS Agents

Reinforcement learning has reshaped large language models through RLHF (Ziegler et al., 2019; Stiennon et al., 2022; Ouyang et al., 2022; Rafailov et al., 2023). More recently, group-based RL methods such as GSPO (Zheng et al., 2025), Dr. GRPO (Liu et al., 2025b), and DAPO (Yu et al., 2026) have achieved strong performance without learning an explicit value function. However, applying RL to information seeking agents remains challenging because long trajectories make outcome-only supervision sparse and noisy.

Recent work has explored denser credit assignment for multi-turn agents. IGPO (Wang et al., 2026) rewards each turn by its marginal gain in correct answer probability, MT PPO (Wei et al., 2025b) uses an LLM judge to score individual steps, and GiGPO (Feng et al., 2026) contrasts actions through anchor state grouping. These methods reduce reward sparsity, but they mainly assign credit through turns, states, or answer probability changes. In web-based information seeking, the value of a tool action depends on the external information it introduces. ICA instead assigns credit through acquired information units and propagates the resulting rewards back to the search and fetch actions that introduced useful sources or content.

### 2.3 Information Seeking Benchmark

The landscape of information-seeking tasks has evolved from basic single-hop queries to complex, multi-hop reasoning, and eventually to highly nonlinear, deep-search challenges. Early single-hop benchmarks, such as NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and SimpleQA-Verified (Haas et al., 2025), primarily consist of questions resolvable through structured queries, simple search operations, or even the model's internal parametric knowledge. Subsequent multi-hop datasets, including Bamboogle (Press et al., 2023), Xbench-DS (Xbench-Team, 2025), and GAIA (Mialon et al., 2024), have significantly elevated the difficulty by introducing multi-step retrieval requirements and domain-specific annotations. These benchmarks rigorously test an agent's capability in deep search and multi-turn tool utilization. At the frontier of this complexity are benchmarks like BrowseComp (Wei et al., 2025a) and Seal-0 (Pham et al., 2026), which represent the most formidable challenges for information-seeking agents. These tasks are characterized by highly nonlinear coupled subproblems and deliberate information masking, introducing substantial uncertainty and search depth that push the boundaries of current agentic reasoning.

## 3 Preliminaries

We consider an IS agent that answers a user query $q \in \mathcal{Q}$ through sequential interactions with external web resources. The agent incrementally acquires external information, updates its interaction history, and produces a final response conditioned on the accumulated evidence.

**State.** At turn $t$, the agent state is defined as the cumulative interaction history

$$s_t = \mathcal{H}_{t-1} = (q, \xi_1, \ldots, \xi_{t-1}), \tag{1}$$

where $q$ is the original user query and each interaction record is defined as

$$\xi_t = (r_t, a_t, o_t). \tag{2}$$

Here, $r_t$ denotes the agent's intermediate reasoning content, $a_t$ denotes the executable action parsed from the agent's output, and $o_t$ denotes the corresponding external observation at turn $t$.

**Action space and interaction protocol.** At turn $t$, the agent chooses an action

$$a_t \in \mathcal{A} = \{\textsc{Search}, \textsc{Fetch}, \textsc{Answer}\}. \quad (3)$$

A SEARCH action issues one or more queries to the web environment, which returns a ranked list of search results

$$o_t^{\text{search}} = \{u_{t,k}\}_{k=1}^{n}, \quad (4)$$

where each result contains a URL, a title, and a short snippet. A FETCH action selects a subset of previously observed URLs, denoted by $L_t$, and retrieves their corresponding contents:

$$o_t^{\text{fetch}} = \{\mathcal{C}(\ell) : \ell \in L_t\}. \quad (5)$$

Here, $\mathcal{C}(\ell)$ denotes the information returned from URL $\ell$, instantiated as either parsed webpage text or a rendered page snapshot.

The interaction history is updated sequentially as

$$\mathcal{H}_t = \mathcal{H}_{t-1} \oplus (r_t, a_t, o_t), \quad (6)$$

where $\oplus$ denotes sequence concatenation. The process terminates when the agent selects ANSWER, after which the final response is generated as

$$a_T = \textsc{Answer}, \qquad \hat{y} \sim \pi_\theta(\cdot \mid \mathcal{H}_{T-1}). \quad (7)$$

## 4 Information-Aware Credit Assignment for Information-Seeking Agents

To address the challenges above, we propose *Information-Aware Credit Assignment (ICA)*, a turn-level credit assignment mechanism for long-horizon information-seeking agents. ICA identifies which action turns are useful for solving the task by examining whether they acquire informative external evidence, and assigns learning signals to those turns instead of uniformly propagating the final outcome reward across the trajectory.

### 4.1 Turn-Level Credit Assignment

Given a prompt $x$, we sample a group of trajectories $G = \{\tau_1, \ldots, \tau_N\}$ from $\pi_{\theta_{\text{old}}}$, where each trajectory receives a binary outcome reward $R_i \in \{0, 1\}$. To address the long-horizon credit assignment challenge, ICA estimates the utility of URLs $U_t^i$ acquired at each action turn $a_t^i$. The URLs are then ranked by their utility scores, and the highest-ranked ones are selected as useful evidence units. Tool turns containing any selected evidence unit receive positive credit, while the final answer turn is supervised by the trajectory-level outcome reward. For failed trajectories, ICA further favors non-terminal exploratory turns over premature answer attempts, encouraging the agent to continue evidence acquisition when the available information is insufficient.

**Atomic-Evidence Counterfactual Contribution** We define an atomic evidence unit $e$ as the minimal identifiable unit of external information that can be acquired and reused by the agent. Let $\mathcal{E}$ denote the union of all atomic evidence units acquired across trajectories in the rollout group. In our implementation, an atomic evidence unit corresponds to a stable URL-level evidence item, specifically an individual result entry returned by SEARCH or a complete webpage snapshot acquired by FETCH.

For each completed trajectory $\tau_i$, let $\mathcal{E}_i \subseteq \mathcal{E}$ denote the set of atomic evidence units in the $i$-th trajectory. Then, the acquisition indicator is defined as

$$I_e^i \;=\; \mathbb{I}[e \in \mathcal{E}_i] \in \{0, 1\}. \quad (8)$$

Each trajectory is associated with a binary outcome $R_i \in \{0, 1\}$ indicating task success.

We estimate the empirical success probability conditioned on acquiring $e$ as

$$P_e^+ = \frac{\sum_{i=1}^{N} I_e^i R_i}{\sum_{i=1}^{N} I_e^i}, \quad (9)$$

and similarly, the success probability conditioned on not acquiring $e$ as

$$P_e^- = \frac{\sum_{i=1}^{N} \left(1 - I_e^i\right) R_i}{\sum_{i=1}^{N} \left(1 - I_e^i\right)}. \quad (10)$$

When the denominator of either estimator is zero, we fall back to the batch-level success rate $P_e^+ = P_e^- = \frac{1}{N}\sum_{i=1}^{N} R_i$ to ensure numerical stability.

The atomic-evidence counterfactual contribution of $e$ is then defined as the difference between these two conditional success probabilities:

$$\Delta_e = P_e^+ - P_e^-. \quad (11)$$

We compute $\Delta_e$ only for groups with mixed outcomes, i.e., groups in which at least one trajectory succeeds and at least one trajectory fails. If all trajectories in a group have the same outcome, the group provides no contrastive signal for tool-level credit assignment, and all tool-turn advantages are set to zero. Importantly, $\Delta_e$ is used only to rank

atomic evidence units, rather than being used directly as a continuous reward. In subsequent ablation studies, we empirically demonstrate a strong correlation between $\Delta_e$ and answer correctness, validating its efficacy as a proxy for information utility.

For each tool type $c \in \{\textsc{Search}, \textsc{Fetch}\}$, let $\mathcal{E}^c$ denote the set of atomic evidence units acquired by tool type $c$ within the group. We rank atomic evidence units according to their counterfactual contribution $\Delta_e$, and select the top positive units:

$$\mathcal{T}^c = \big\{ e \in \mathcal{E}^c \;\big|\; \mathrm{rank}(e) \le k_c \wedge \Delta_e > 0 \big\}, \quad (12)$$

where

$$k_c = \max\Big(1,\ \lfloor \rho |\mathcal{E}^c| \rfloor\Big). \quad (13)$$

Here, $\rho \in (0, 1)$ is a hyperparameter governing the top fraction of retrieved URLs to be designated as high-utility evidence, and $\mathrm{rank}(e)$ denotes the descending rank order of $e$ based on its counterfactual contribution $\Delta_e$.

**Action-Turn Reward** To decouple search quality from the final outcome, we assign distinct rewards to intermediate tool turns and the terminal answer turn. For any intermediate action turn $t$ invoking a tool of type $c_t^i \in \{\textsc{Search}, \textsc{Fetch}\}$, we evaluate its acquired URL set $\mathcal{E}_t^i$. Instead of relying on continuous counterfactual scores that easily collapse under small group sizes, we assign a binary reward $r_t^i$:

$$r_t^i = \mathbb{I}\big[\mathcal{E}_t^i \cap \mathcal{T}^{c_t^i} \neq \emptyset\big]. \quad (14)$$

This means the tool turn is explicitly rewarded with 1 if it successfully captures at least one high-utility evidence unit from $\mathcal{T}^{c_t^{(i)}}$, and 0 otherwise. Conversely, the terminal answer turn is evaluated via an asymmetric shaped reward $\tilde{R}^{(n)}$:

$$\tilde{R}_i = \begin{cases} 1.0, & \text{if } R_i = 1, \\ -1.0 + \sqrt{\frac{|\tau_i|}{L_{\max}}}, & \text{if } R_i = 0. \end{cases} \quad (15)$$

Here, $R_i \in \{0, 1\}$ indicates the correctness of the final answer. This budget-aware shaping mechanism scales down the penalty based on the consumed token count $|\tau_i|$ relative to the maximum context window $L_{\max}$, thereby protecting thorough search exploration from being heavily penalized upon failure.

**Turn-Level Advantage Assignment** To explicitly decouple search-quality supervision from final-answer outcomes, we assign advantages by centering the rewards strictly within their respective scopes. Each token generated at turn $t$ in trajectory $\tau_i$ inherits the turn-level advantage $A_t^i$:

$$A_t^i = \begin{cases} \frac{r_t^i - \bar{r}}{\sigma_r}, & a_t^i \in \{\textsc{Search}, \textsc{Fetch}\}, \\ \frac{\tilde{R}_i - \bar{R}}{\sigma_R}, & a_t^i \in \{\textsc{Answer}\}. \end{cases} \quad (16)$$

Here, the baselines $\bar{r}$ and $\bar{R}$ are the group-level mean rewards for the specific tool type $c_t^i$ and the final answer, respectively, while $\sigma_r$ and $\sigma_R$ denote their corresponding standard deviations. Specifically, for the intermediate tool calls, we compute $\bar{r} = \frac{1}{|\mathcal{S}^c|} \sum_{(i,t) \in \mathcal{S}^c} r_i^t$, where $\mathcal{S}^c$ denotes the set of all tool turns of type $c$ within the group.

**ICA-GSPO Optimization** We optimize the policy using a turn-level GSPO objective. Building upon the trajectories and advantages established in Sec. 4.1, action turns receive information-acquisition advantages, and the final answer turn receives the outcome-level advantage.

Unlike standard GSPO which computes a single importance ratio over the entire trajectory, ICA-GSPO computes this ratio independently within each turn, inspired by recent advances in turn-level importance sampling (Li et al., 2026). This design aligns with the natural multi-turn structure of the agent and prevents large policy shifts in a single token from contaminating the trajectory-level estimate. For the $t$-th turn of trajectory $\tau_i$, all tokens share a common geometric-mean importance ratio $s_t^i(\theta)$:

$$s_t^i(\theta) = \exp\left( \frac{1}{|\tau_t^i|} \sum_{l \in \tau_t^i} \log \frac{\pi_\theta(y_l \mid \mathcal{H}_{t-1}^i, y_{<l})}{\pi_{\theta_{\text{old}}}(y_l \mid \mathcal{H}_{t-1}^i, y_{<l})} \right). \quad (17)$$

Using this turn-level ratio and the normalized advantage $A_t^i$, we define the turn-level clipped GSPO surrogate objective as

$$\mathcal{J}_t^i(\theta) = \min\Big\{ s_t^i(\theta) A_t^i, \\ \mathrm{clip}\big(s_t^i(\theta), 1-\epsilon, 1+\epsilon\big) A_t^i \Big\}. \quad (18)$$

To prevent the critical learning signal of the single answer turn from being diluted by a potentially large number of preceding tool turns, we aggregate the surrogate objectives of the non-answer turns

and the final answer turn using a fixed weighting factor $w$:

$$\mathcal{J}_i(\theta) = w\, \mathcal{J}^i_{T_i}(\theta) + (1-w)\frac{1}{T_i - 1}\sum_{k=1}^{T_i-1} \mathcal{J}^i_k(\theta). \tag{19}$$

The overall objective is then obtained by taking the expectation of the trajectory-level surrogate objective over trajectories sampled from the old policy:

$$\mathcal{J}_{\text{ICA-GSPO}}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}}\left[\mathcal{J}(\theta;\tau)\right]. \tag{20}$$

### 4.2 External Observations

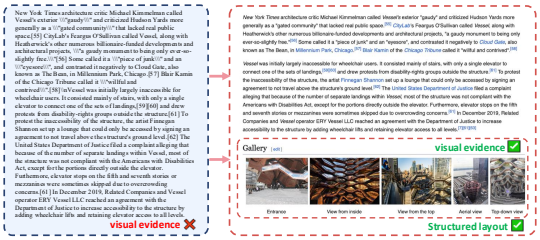
Existing approaches commonly obtain web content in text-only form, either by summarizing webpages or retrieving a subset of passages via RAG-style pipelines (Jin et al., 2025). Such representations introduce irreversible information loss at acquisition time, break cross-trajectory consistency for the same information source, and inject additional model bias through learned summarization or retrieval components. These properties are fundamentally misaligned with ICA, which estimates the utility of external information units based on their aggregate contribution to successful outcomes across trajectories.

To address this issue, we model external observations as webpage snapshots rather than summarized or retrieved text. A snapshot preserves the full visible content of a webpage, including textual elements, layout structure, tables, and images, and is treated as a persistent external information unit in the agent's memory. Snapshots are returned directly by tool invocations and consumed by the agent without intermediate filtering or compression, allowing the agent to determine which information becomes relevant during downstream reasoning, as illustrated in Figure 2. This representation provides a stable and consistent basis for identifying recurring information units across trajectories.

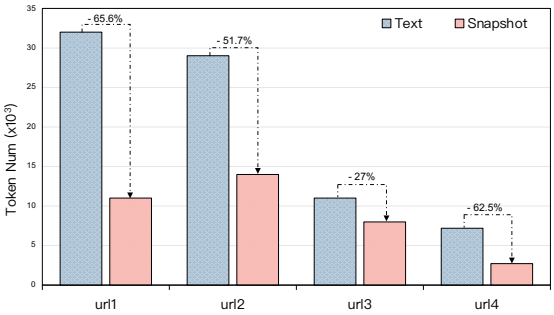
## 5 Experiments

### 5.1 Main Results

This section presents the main empirical results of our approach. We primarily evaluate Information-Aware Credit Assignment on a suite of long-horizon **information-seeking benchmarks**, comparing against both proprietary and open-source agentic systems to assess performance in dynamic web environments that require iterative search, evidence acquisition, and multi-step reasoning.




(a) Full Text & Snapshot-Based Representation




(b) Token Budget Comparison

Figure 2: Text vs Visual snapshots. **(a)** Plain text extracted from webpages is often cluttered and lacks visual structure, while snapshots preserve layout and non-textual cues. **(b)** Token count comparison across four URLs, showing that snapshots reduce token usage by 27.0% – 65.6% compared to parsed text.

#### 5.1.1 Performance on Information Seeking Benchmarks

Table 1 compares our approach with proprietary and open-source agentic systems on long-horizon information-seeking benchmarks. Overall, **ICA** demonstrates strong and consistent performance across all evaluated tasks, validating the effectiveness of Information-Aware Credit Assignment in complex, multi-step web environments.

Compared with proprietary systems, ICA-30B-A3B remains competitive on challenging benchmarks such as BrowseComp and Seal-0, despite using a significantly smaller model size and a fully open training pipeline.

#### 5.1.2 Ablation Study

Table 2 isolates the effects of (i) the fetch representation used in SFT (text parsing vs. Snapshot visual grounding) and (ii) the RL optimization scheme (comparing GSPO with ICA).

**Comparison of Snapshot and Text in SFT.** Across both backbones, adopting snapshot-based fetches in SFT yields consistent improvements

Table 1: Comparison of different models on information-seeking benchmarks. We report the Pass@1 accuracy on BrowseComp, GAIA, Xbench-DS, and Seal-0, while "snap." refers to visual snapshot-based fetching. Results marked with * are taken from existing studies. Best open-source results are highlighted in **bold**.

| Model / Framework | Tools | BrowseComp | GAIA | Xbench-DS | Seal-0 |
|---|---|---|---|---|---|
| ***Proprietary Agents*** | | | | | |
| Claude-4-Sonnet* | – | 12.2 | 68.3 | 64.6 | – |
| OpenAI-o3* | – | 49.7 | 70.5 | 66.7 | 18.9 |
| OpenAI DeepResearch* | – | 51.5 | 67.4 | – | – |
| ***Open-Source Agents (<15B)*** | | | | | |
| WebExplorer-8B* | Search & Fetch (text) | 15.7 | 50.0 | 53.7 | – |
| WebSailor-7B* | Search & Fetch (text) | 6.7 | – | 34.3 | – |
| DeepDive-9B* | Search & Fetch (text) | 6.3 | – | 38.0 | 12.2 |
| MiroThinker-14B-DPO-v0.1* | Search & Fetch (text) | 9.0 | – | 30.0 | – |
| Qwen3-14B* | Search & Fetch (text) | 1.0 | – | 20.0 | – |
| InfoAgent* | Search & Fetch (text) | 15.3 | – | 40.4 | – |
| Qwen3-VL-8B-ICA (Ours) | Search & Fetch (snap.) | **25.0** | **69.9** | **71.0** | **25.2** |
| ***Open-Source Agents (>15B)*** | | | | | |
| ASearcher-Web-32B* | Search & Fetch (text) | 5.2 | 52.8 | 42.1 | – |
| MiroThinker-32B-DPO-v0.2* | Search & Fetch (text) | 13.0 | 64.1 | – | – |
| Kimi-K2-Instruct-1T* | Search & Fetch (text) | 14.1 | 57.7 | 50.0 | – |
| WebDancer-QwQ-32B* | Search & Fetch (text) | 3.8 | 51.5 | 38.3 | – |
| WebSailor-32B* | Search & Fetch (text) | 10.5 | 53.2 | 53.3 | 21.3 |
| DeepDive-32B* | Search & Fetch (text) | 15.3 | – | 51.8 | 25.5 |
| C-GSPO* | Search & Fetch (text) | 24.8 | 56.3 | 57.7 | – |
| WebShaper-QwQ-32B* | Search & Fetch (text) | – | 53.3 | 35.0 | – |
| WebLeaper* | Search & Fetch (text) | 22.7 | 69.9 | 62.3 | **35.1** |
| Qwen3-VL-30B-A3B-ICA (Ours) | Search & Fetch (snap.) | **26.1** | **72.8** | **76.0** | 27.0 |

Table 2: Ablation study of model components. **BC-100** denotes a subset of 100 queries randomly sampled from BrowseComp, and **XDS** denotes Xbench-DS. All results are reported as Pass@1 accuracy.

| Method | BC-100 | GAIA | XDS | Seal-0 |
|---|---|---|---|---|
| **Baseline: Qwen3-VL-8B-Thinking** | | | | |
| Baseline - Text | 1.0 | 29.1 | 39.0 | 7.2 |
| SFT - Text | 11.0 (+10.0) | 54.3 (+25.2) | 55.0 (+16.0) | 20.7 (+13.5) |
| SFT - Snap. | 18.0 (+17.0) | 56.4 (+27.3) | 66.0 (+27.0) | 23.4 (+16.2) |
| GSPO - Snap. | 22.0 (+21.0) | 60.2 (+31.1) | 68.0 (+29.0) | 24.3 (+17.1) |
| ICA - Snap. | 25.0 (+24.0) | 69.9 (+40.8) | 71.0 (+32.0) | 25.2 (+18.0) |

over textual fetches, with gains concentrated on benchmarks that are more sensitive to layout and presentation (e.g., GAIA and XDS). This indicates that snapshot observations better preserve layout-dependent semantics while reducing artifacts introduced by heuristic webpage parsing and text linearization, thereby enabling more reliable evidence localization during multi-step retrieval.

**Advantages of ICA in RL Optimization.** With snapshot retrieval fixed as the fetch policy during RL, ICA consistently outperforms vanilla GSPO on all benchmarks and both model scales. The improvements are particularly salient on long-horizon, high-noise tasks, supporting the claim that information-aware post-hoc credit assignment provides a denser and more targeted training signal by attributing outcome rewards to the retrieval turns that introduce high-utility evidence.

### 5.1.3 Analysis of ICA

To further examine potential confounds and better understand the behavior of ICA, we conduct additional diagnostic experiments focusing on the quality of selected atomic evidence under controlled trajectory lengths.

Table 3: Evaluation under strict context constraints (128k context limit, 16k tool response truncation threshold). We report Pass@1 accuracy with the context overflow rate in parentheses. **XDS** denotes Xbench-DS.

| Method | BC-100 | GAIA | XDS | Seal-0 |
|---|---|---|---|---|
| SFT - Text | 11.0 (36.0) | 54.3 (15.5) | 55.0 (6.0) | 20.7 (20.0) |
| SFT - Snap. | **18.0 (0.0)** | **56.4 (0.0)** | **66.0 (0.0)** | **23.4 (0.0)** |

**Length-controlled sufficiency.**
To isolate the effect of trajectory length, we partition trajectories into five buckets based on the number of interaction steps: 0–19, 20–39, 40–59, 60–79, and 80–99. Within each bucket, we evaluate the sufficiency of ICA-selected evidence using a dynamic *top-*$p\%$ protocol (where $p \in \{10, 15, 20\}$): we retain the top $p\%$ highest-scoring atomic evidence units according to ICA and prompt the model to answer using only these units. We compare this against two baselines:
(i) randomly sampled evidence of the same proportion (*random-*$p\%$)
(ii) the lowest-scoring evidence under ICA (*tail-*$p\%$).

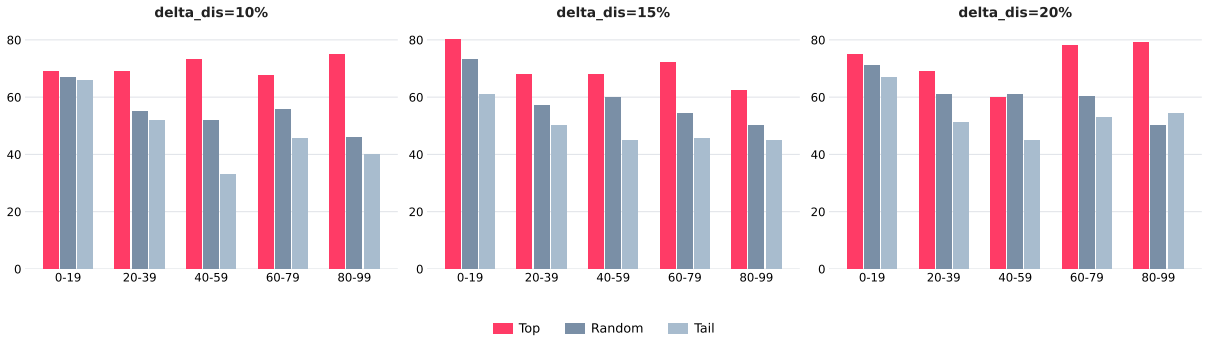



Figure 3: Top-$p\%$ ICA evidence (red) consistently outperforms random and tail baselines across all trajectory lengths and thresholds ($p \in \{10, 15, 20\}$).

As shown in Figure 3, the top-$p\%$ ICA evidence consistently achieves the highest answer accuracy across almost all trajectory lengths and filtering thresholds, with detailed values reported in Appendix Table 4. On average across these subsets, the ICA-selected evidence reaches an accuracy of 71.0%, substantially outperforming both the random (58.2%) and tail (50.2%) baselines. These results indicate that ICA reliably assigns higher scores to atomic evidence that is more likely to be sufficient for answering the question. Importantly, this advantage holds across all trajectory buckets and selection ratios, suggesting that the effectiveness of ICA is not driven by a bias toward longer trajectories, but rather by its robust ability to identify answer-supporting evidence regardless of the search length.

**Token Efficiency and Context Overflow.** To investigate the context efficiency of different webpage representation modalities, we conduct a controlled evaluation under strict context length and token budget constraints. Specifically, we cap the maximum context window of the 8B model to 128k tokens. Furthermore, the token budget for each individual tool response is strictly bounded within a 16k window; any returning content exceeding this threshold is forcefully truncated.

We report the Pass@1 accuracy alongside the context overflow rate (i.e., the percentage of trajectories that trigger context truncation) across four benchmarks in Table 3. Under identical context and tool response constraints, the Snapshot modality (SFT - Snapshot) consistently outperforms the raw text baseline (SFT - Text) in both task efficacy and token efficiency. Crucially, while the text modality suffers from severe context overflow, reaching 36.0% on BrowseComp and 20.0% on Seal-0, the snapshot modality has an overflow rate of only 0.0% across all evaluated datasets. This drastically lower token footprint translates directly to substantial performance gains, boosting Pass@1 accuracy by 7.0% on BrowseComp and 11.0% on Xbench-DS. Notably, on GAIA, the snapshot modality improves Pass@1 accuracy from 54.3% to 56.4% while eliminating context overflow, whereas the text baseline still incurs a 15.5% overflow rate. These results demonstrate that Snapshot web representations are inherently more compact than raw HTML text, effectively mitigating context bloat and enabling robust navigation within constrained budgets.

## 6 Conclusion

This work addresses two compounding bottlenecks in long-horizon open-web information seeking: the fidelity loss of text-derived webpage representations and sparse credit assignment over extended retrieval trajectories. By representing fetched pages as rendered snapshots, our framework provides stable, identifiable evidence units that can be reliably compared across trajectories, enabling Information-Aware Credit Assignment (ICA) to estimate the counterfactual utility of each unit via posterior success association and propagate dense turn-level rewards to upstream retrieval decisions. Experiments on BrowseComp, GAIA, Xbench-DS, and Seal-0 demonstrate consistent gains from both snapshot grounding and ICA, with our 8B model surpassing all open-source agents under 15B. These results suggest that performing credit assignment at the granularity of acquired evidence, rather than relying solely on terminal outcome signals, is a practical and effective strategy for training long-horizon web agents.

## Limitations

While our proposed method demonstrates strong performance on long-horizon information-seeking tasks, we acknowledge two primary limitations in our current design.

First, regarding the webpage processing mechanism, our framework relies on uncompressed visual representations. When the agent fetches a webpage, the raw rendered snapshot is directly fed into the Vision Transformer (ViT) and converted into visual tokens. We currently do not apply any intermediate visual compression, spatial pooling, or salient region filtering prior to tokenization. Although this straightforward approach preserves the original visual layout and fine-grained multi-modal cues of the web environment, it inevitably introduces considerable token redundancy and computational overhead, particularly when processing visually dense or exceedingly long webpages.

Second, our system currently lacks a sophisticated context management strategy to handle unbounded exploration. As the agent continuously executes tools and accumulates evidence across multiple turns, the sequence length grows rapidly. Without dynamic context reduction techniques, the accumulated tokens can easily exceed the model's maximum context limit during continuous exploration and evidence accumulation, particularly when historical summarization, token eviction, or memory-based sliding windows are not applied. In such cases of context overflow, the trajectory must be forcefully truncated, which may lead to the loss of early-retrieved crucial evidence and constrain the agent's capability on tasks that require extensively deep search steps. Addressing these efficiency bottlenecks through visual token compression and adaptive memory management remains an important direction for future work.

## References


Anthropic. 2025. Introducing Claude 4.

Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, Chen Sun, Han Hou, Hui Yang, James Pan, Jianan Lou, Jiayi Mao, Jizheng Liu, Jinpeng Li, Kangyi Liu, and 14 others. 2025. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *Preprint*, arXiv:2506.13651.

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

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. 2025. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl. *Preprint*, arXiv:2508.07976.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. 2025. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*.

Guanzhong He, Zhen Yang, Jinxin Liu, Bin Xu, Lei Hou, and Juanzi Li. 2026. Webseer: Training deeper search agents through reinforcement learning with self-reflection. In *The Fourteenth International Conference on Learning Representations*.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Chenliang Li, Adel Elmahdy, Alex Boyd, Zhongruo Wang, Siliang Zeng, Alfredo Garcia, Parminder Bhatia, Taha Kass-Hout, Cao Xiao, and Mingyi Hong. 2026. Stabilizing off-policy training for long-horizon llm agent via turn-level importance sampling and clipping-triggered normalization. *Preprint*, arXiv:2511.20718.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025a. Websailor: Navigating super-human reasoning for web agent. *Preprint*, arXiv:2507.02592.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025b. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5420–5438, Suzhou, China. Association for Computational Linguistics.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025c. Webthinker: Empowering large reasoning models with deep research capability. In *Advances in Neural Information Processing Systems*, volume 38, pages 120091–120131. Curran Associates, Inc.

Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhu Chen, Pengyu Zhao, and Junxian He. 2025a. Webexplorer: Explore and evolve for training long-horizon web agents. *Preprint*, arXiv:2509.06501.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*.

Rui Lu, Zhenyu Hou, Zihan Wang, Hanchen Zhang, Xiao Liu, Yujiang Li, Shi Feng, Jie Tang, and Yuxiao Dong. 2025. Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl. *Preprint*, arXiv:2509.10446.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*.

OpenAI. 2025a. Deep research system card. OpenAI Official System Card.

OpenAI. 2025b. OpenAI o3 and o4-mini. OpenAI Official Announcement.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Thinh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. 2026. Sealqa: Raising the bar for reasoning in search-augmented language models. *Preprint*, arXiv:2506.01062.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *Preprint*, arXiv:2503.05592.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Zhengwei Tao, Haiyang Shen, Baixuan Li, Wenbiao Yin, Jialong Wu, Kuan Li, Zhongwang Zhang, Huifeng Yin, Rui Ye, Liwen Zhang, and 1 others. 2025. Webleaper: Empowering efficiency and efficacy in webagent via enabling info-rich seeking. *Preprint*, arXiv:2510.24697.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Pu Wu, Junkai Zhang, Baixuan Li, Haiyang SHEN, Kuan Li, Liwen Zhang, Xinyu Wang, Wentao Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2026. Webshaper: Agentically data synthesizing via information-seeking formalization. In *The Fourteenth International Conference on Learning Representations*.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, and 181 others. 2026a. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, Wenhan Dou, Yue Deng, Yunjie Fu, Junqi Ge, Chenxia Han, Tammy Huang, Zhenhang Huang, Jerry Jiao, Shilei Jiang, and 36 others. 2026b. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *Preprint*, arXiv:2511.11793.

Tongyi DeepResearch Team, 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, and 38 others. 2026c. Tongyi deepresearch technical report. *Preprint*, arXiv:2510.24701.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. 2026. Information gain-based policy optimization: A simple and effective approach for multi-turn

search agents. In *The Fourteenth International Conference on Learning Representations*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025a. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. 2025b. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *Preprint*, arXiv:2505.11821.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, Ding-Chu Zhang, Zekun Xi, Xiangru Tang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2026. Webdancer: Towards autonomous information seeking agency. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. WebWalker: Benchmarking LLMs in web traversal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10290–10305, Vienna, Austria. Association for Computational Linguistics.

xAI. 2025. Grok 3 beta — the age of reasoning agents. Website. Accessed: 2026-01-29.

Xbench-Team. 2025. Xbench-deepsearch.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2026. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Gongrui Zhang, Jialiang Zhu, Ruiqi Yang, Kai Qiu, Miaosen Zhang, Zhirong Wu, Qi Dai, Bei Liu, Chong Luo, Zhengyuan Yang, Linjie Li, Lijuan Wang, Weizhu Chen, Yuan Zhang, Xin Li, Zhaoyi Liu, Xin Geng, and Baining Guo. 2025. Infoagent: Advancing autonomous information-seeking agents. *Preprint*, arXiv:2509.25189.

Jiajie Zhang, Xin Lv, Ling Feng, Lei Hou, and Juanzi Li. 2026. Chaining the evidence: Robust reinforcement learning for deep search agents with citation-aware rubric rewards. *Preprint*, arXiv:2601.06021.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

# A Experimental Details

We will provide a detailed description of the algorithmic workflow and the specific experimental setup in this section.

## A.1 Experiment Setup

**Environment Implementation** For the environment implementation, we utilize the Serper API[1] for the `search` tool, while the `fetch` tool is instantiated in two modalities: a text-based baseline using Trafilatura[2] to retrieve parsed content, and our proposed snapshot-based setting using Playwright[3] to capture full-page visual observations.

**Benchmarks & Metrics** We evaluate our method on a suite of long-horizon information-seeking benchmarks that require iterative search, evidence acquisition, and multi-step reasoning in dynamic web environments. Specifically, we adopt BrowseComp (Wei et al., 2025a), GAIA (Mialon et al., 2024), Xbench-DS (Xbench-Team, 2025), and Seal-0 (Pham et al., 2026) to assess agentic information-seeking capabilities under varying task complexities and environmental dynamics. For GAIA, we evaluate on the 103 examples from the text-only validation subset (Li et al., 2025b), following standard evaluation protocols. For Xbench-DS, we utilize the more mature 2505 version release for evaluation. Across all benchmarks, we report Pass@1 as the primary metric, obtained via LLM-as-a-judge evaluation.

**Baselines** We compare against a diverse set of strong baselines spanning both proprietary and open-source information-seeking agents. These include proprietary systems such as Claude-4-Sonnet (Anthropic, 2025), OpenAI-o3 (OpenAI, 2025b), and DeepResearch (OpenAI, 2025a), as well as leading open-source methods including ASearcher (Gao et al., 2025), DeepDive (Lu et al., 2025), MiroThinker (Team et al., 2026b), InfoAgent (Zhang et al., 2025), Kimi-K2 (Team et al., 2026a), WebExplorer (Liu et al., 2025a), WebDancer (Wu et al., 2026), WebSailor (Li et al., 2025a), WebShaper (Tao et al., 2026), WebLeaper (Tao et al., 2025), and C-GSPO (Zhang et al., 2026).

**Training Settings and Hyperparameter Details** Unless otherwise specified, all models are trained based on Qwen3-VL-8B-Thinking and Qwen3-VL-30B-A3B-Thinking. During rollout generation, we use a sampling temperature of 0.6 and top-p nucleus sampling with p = 0.95. We set the fixed weighting factor to 0.3 and the top-p evidence selection ratio to 10%.

## A.2 Compute Budget

All models were trained for 50 steps on NVIDIA H800 GPUs. The computational cost for the 8B model was approximately 443 GPU-hours. For the 30B-A3B model, the computational cost increased to approximately 534 GPU-hours.

## A.3 Tools Description

**RAG Text Tool** To align with other models, facilitate distillation, and provide a robust cold start, we implement an intelligent retrieval and reranking module, **RAGTool**. The module performs content acquisition, text normalization/chunking, and relevance scoring in a unified workflow. For content acquisition, we employ a high-throughput crawler based on the `httpx` asynchronous framework, supporting up to **32 concurrent** requests to efficiently gather candidate pages. For access-restricted or unstable webpages (e.g., dynamic rendering, interstitials, or anti-bot behaviors), we incorporate the JINA API as a fallback mechanism to improve the robustness and coverage of raw text retrieval while maintaining a consistent downstream interface. Retrieved HTML is then cleaned and normalized (removing boilerplate and redundant markup) and segmented under a token-budget constraint to fit embedding/reranking context limitations. Concretely, the base chunk size is set to **256 tokens**, and a lightweight `naive_merge` strategy merges under-length segments with adjacent neighbors to reduce semantic fragmentation at chunk boundaries. For relevance estimation, we adopt **BAAI/bge-reranker-v2-m3**, which provides strong

[1] https://serper.dev
[2] https://github.com/adbar/trafilatura
[3] https://playwright.dev

semantic matching and cross-lingual retrieval capabilities, to compute query–chunk relevance scores. Candidate chunks are ranked in descending order, and only the **Top-$K$** high-confidence segments are retained ($K = 10$). Finally, the concatenated evidence is dynamically truncated to a maximum context budget (default **2048 tokens**) so that limited downstream context is dominated by high-salience signals, thereby improving the information density for subsequent reasoning.

**Full-Text Tool** To serve as a direct text-based baseline for our text and snapshot comparative experiments, we introduce a straightforward HTML parsing module that bypasses semantic reranking. Following the identical high-throughput content acquisition pipeline (leveraging asynchronous `httpx` with the JINA API fallback), this tool directly processes the retrieved HTML to extract the complete raw text. After stripping standard boilerplate and redundant structural markup, the entire textual content is retained in its original linear order. To ensure a fair comparison with the visual snapshot modality regarding token consumption, the extracted full text is simply truncated to a maximum token limit that precisely matches the snapshot's context window size. This naive extraction and truncation mechanism deliberately preserves the original flow of the webpage, allowing us to explicitly evaluate the impact of context overflow and token efficiency when comparing uncompressed text representations against multi-modal visual snapshots.

**Snapshot Tool** To mitigate resolution constraints and context loss when vision–language models process long webpages, we develop a visual perception component, **SnapshotTool**, that integrates high-fidelity rendering, adaptive slicing, and visual optimization within a single pipeline. The system is implemented with Playwright and maintains a **browser pool of 8 Chromium instances**, supporting up to **32 concurrent** rendering requests for scalable batch processing. During rendering, an auto-scrolling procedure is applied to trigger lazy-loaded elements and ensure comprehensive visual capture, while a lightweight denoising routine suppresses non-content overlays (e.g., cookie banners, modal dialogs, and popups) through targeted DOM manipulation and style injection to provide clean visual inputs. For long pages, we enforce a hard maximum rendering height of **20,000 pixels**, which covers the primary information region of most websites while bounding worst-case computation. Within this range, a sliding-window strategy produces a sequence of serialized screenshot slices with slice height **4,480 pixels** and vertical overlap **112 pixels**, reducing boundary truncation artifacts (e.g., text lines, figures, and tables split across slice edges) and preserving visual semantic continuity across slices. To reduce downstream VLM input overhead without sacrificing readability, each slice is uniformly downsampled by a **0.7**$\times$ factor using **LANCZOS** interpolation. When local rendering fails, the module falls back to the **Jina API** and applies the same slicing and compression logic, ensuring consistent output structure and comparable visual characteristics across rendering pathways.

### A.4 Data Curation

**Supervised Finetuning Data.** To equip the model with foundational web-interaction skills and establish a stable initialization for reinforcement learning, we construct a hybrid supervised fine-tuning (SFT) dataset consisting entirely of high-difficulty, multi-step reasoning problems. This complex dataset comprises two complementary subsets distilled from different observation modalities.

The first subset leverages an automated framework that integrates hierarchical knowledge extraction on Wikipedia with adaptive web exploration to synthesize a pool of sophisticated QA pairs. Trajectories for these complex problems are distilled using DeepSeek under a *text-based* fetching modality. To ensure a fair comparison regarding token budgets, we apply a goal-guided reranking strategy to extract the most relevant textual chunks, intentionally establishing a remarkably strong text baseline.

The second subset focuses on visual web navigation for similarly challenging tasks, utilizing diverse, high-difficulty information-seeking questions sampled from ASearcher (Gao et al., 2025). For these queries, we distill trajectories using GLM 4.6 under a *snapshot-based* fetching modality, which returns webpage snapshots to explicitly preserve layout and structural cues. By combining these two distillation sources, the SFT phase trains the model to tackle complex web-search mechanics while becoming fluent in both highly-optimized textual and visual web representations.

**Reinforcement Learning Data.** To drive policy optimization during the RL phase, the agent requires challenging environments that demand deep, autonomous exploration. Rather than relying on generic queries, we curate a specialized dataset of 0.4k high-difficulty, multi-layered questions. These queries are explicitly selected from the complex problem pool generated by our aforementioned automated synthesis framework. By utilizing these carefully constructed long-tail questions as the initial prompts, we encourage the RL agent to engage in multi-turn exploration and complex evidence alignment without relying on step-by-step expert demonstrations.

Table 4: Answer accuracy (%) across varying trajectory lengths (interaction steps). The evaluation is conducted under different evidence filtering thresholds ($p \in \{10, 15, 20\}$). The best results within each threshold setting are highlighted in bold.

| Threshold | Method | Trajectory Length (Steps) | | | | | Avg. |
|---|---|---|---|---|---|---|---|
| | | 0–19 | 20–39 | 40–59 | 60–79 | 80–99 | |
| $p = 10\%$ | Top-10% | **69.0** | **69.0** | **73.0** | **67.7** | **75.0** | **70.7** |
| | Random-10% | 67.0 | 55.0 | 52.0 | 55.9 | 45.8 | 55.1 |
| | Tail-10% | 66.0 | 52.0 | 33.0 | 45.6 | 40.0 | 47.3 |
| $p = 15\%$ | Top-15% | **80.0** | **68.0** | **68.0** | **72.1** | **62.5** | **70.1** |
| | Random-15% | 73.0 | 57.0 | 60.0 | 54.4 | 50.0 | 58.9 |
| | Tail-15% | 61.0 | 50.0 | 45.0 | 45.6 | 45.2 | 49.3 |
| $p = 20\%$ | Top-20% | **75.0** | **69.0** | **61.0** | **77.9** | **79.2** | **72.2** |
| | Random-20% | 71.0 | 61.0 | 60.0 | 60.3 | 50.0 | 60.7 |
| | Tail-20% | 67.0 | 51.0 | 45.0 | 52.9 | 54.2 | 54.0 |

Table 5: Results on BrowseComp, GAIA, Xbench-DS, Seal-0. The results are evaluated with LLM-as-Judge. For baselines, we run the evaluation for 4 seeds and report **Pass@4**, which reflects the model's upper bound. Results marked with * are taken from existing studies.

| Method | Fetch Modality | BrowseComp | GAIA | Xbench-DS | Seal-0 |
|---|---|---|---|---|---|
| QwQ-32B Direct Gen.* | Text | – | 31.1 | 23.0 | – |
| Search-o1 (QwQ)* | Text | – | 67.0 | 65.0 | – |
| Search-R1-32B* | Text | – | 43.7 | 37.0 | – |
| WebThinker-QwQ* | Text | – | 57.3 | 52.0 | – |
| Simple DS-QwQ* | Text | – | 64.1 | 61.0 | – |
| WebDancer-QwQ* | Text | – | 61.2 | 68.0 | – |
| ASearcher-Web-QwQ-v2* | Text | – | 70.1 | 68.0 | – |
| Qwen3-VL-8B-ICA (Ours) | Snapshot | 39.9 | 86.4 | 87.0 | 49.3 |
| Qwen3-VL-30B-A3B-ICA (Ours) | Snapshot | **45.0** | **89.3** | **89.0** | **52.3** |

To facilitate reproducibility, we provide the pseudocode of our ICA in Algorithm 1.

## B Additional Experimental Results

Table 5 summarizes results on BrowseComp, GAIA, XBench-DS, and Seal-0 under an LLM-as-Judge evaluation protocol. These benchmarks are designed to stress web agents' ability to browse real pages, follow long-horizon reasoning chains, and ground answers on external evidence rather than relying on the LLM's internal knowledge. Among the text-fetching baselines, ASearcher-Web-QwQ-v2 is the strongest on GAIA (70.1) and XBench-DS (68.0). In contrast, our snapshot-based ICA consistently delivers the best performance across all reported datasets: ICA-8B already reaches 86.4 on GAIA and 87.0 on XBench-DS, and scaling to ICA-30B-A3B further boosts results to 89.3 and 89.0, surpassing the best text baseline by +19.2 and +21.0 points, respectively. Moreover, ICA-30B-A3B achieves 45.0 on BrowseComp and 52.3 on Seal-0, demonstrating robust gains on challenging browsing-centric tasks. For baselines, we run the official implementations with four random seeds and report Pass@4, which reflects the model's upper-bound performance.

**Algorithm 1** ICA-GSPO

**Require:** Policy $\pi_\theta$, question $q$, high-utility evidence sets $\{\mathcal{T}^c\}$, maximum context length $L_{\max}$
**Ensure:** Updated policy parameters $\theta$

Sample a group of trajectories $\{\tau_i\}_{i=1}^N \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$
Evaluate final answers to obtain correctness outcomes $\{R_i\}_{i=1}^N \in \{0, 1\}$
**for** each trajectory $\tau_i$ **do**
  *1. Compute budget-aware shaped reward for the final answer*
  **if** $R_i = 1$ **then**
    $\tilde{R}_i \leftarrow 1.0$
  **else**
    $\tilde{R}_i \leftarrow -1.0 + \sqrt{|\tau_i|/L_{\max}}$
  **end if**
  *2. Compute binary rewards for intermediate tool calls*
  **for** each tool-invoking turn $t$ with tool type $c_t^i$ **do**
    Extract acquired evidence/URL set $\mathcal{E}_i$
    $r_t^i \leftarrow \mathbb{I}\big[\mathcal{E}_i \cap \mathcal{T}^{c_t^i} \neq \emptyset\big]$
  **end for**
**end for**
*3. Compute group-level normalization statistics*
Compute $(\bar{R}, \sigma_R)$ over final shaped rewards $\{\tilde{R}_i\}_{i=1}^N$
Compute $(\bar{r}, \sigma^r)$ over $\{r_t^i\}$ independently for each tool type $c \in \{\text{SEARCH}, \text{FETCH}\}$
**for** each trajectory $\tau_i$ **do**
  *# 4. Assign turn-level decoupled advantages*
  **for** each turn $t$ in $\tau_i$ **do**
    **if** $t$ is a tool-call turn of type $c_t^i$ **then**
      $A_t^i \leftarrow \frac{r_t^i - \bar{r}}{\sigma_r + \epsilon}$
    **else if** $t$ is the final answer turn **then**
      $A_t^i \leftarrow \frac{\tilde{R}_i - \bar{R}}{\sigma_R + \epsilon}$
    **end if**
  **end for**
**end for**
Update $\theta$ using policy gradient/GSPO surrogate loss with advantages $A_t^i$

## C Judge Prompt

**Judge prompt**

You grade whether a model's answer matches the standard answer for a benchmark question.

Inputs:

- [Question]: the task.
- [Standard Answer]: the reference (treated as correct).
- [Model_answer]: the model's full or post-processed output (may include reasoning, citations, or how it found the answer).

Your job: decide if the substantive content of [Model_answer] matches [Standard Answer] for what the question asks.

Output 1 (correct) when:

- The model's final factual claim matches the standard answer, even if phrased differently, in different order, or with extra units/formatting (e.g. 0.1777 $m^3$ vs 0.1777, "Annie Levin" vs "**Annie Levin**").
- Extra reasoning, search steps, quotes, footnote markers like [^0^], or URLs are OK if the resolved answer still matches.

Output 0 (incorrect) when:

- The model gives a different number/name/date/entity, contradicts the standard answer, or is off-topic.
- The model clearly refuses or says it cannot answer.
- The model only restates the question or only copies the standard answer without committing to an answer (no substantive solution).

Important: Describing that information was found via search, browsing, or tools does NOT by itself make the answer wrong. Grade the correctness of the answer content, not the presence of tool narration.

If they are consistent, Judgement is 1; if they are different, Judgement is 0.

Just output Judgement and don't output anything else.