SOCAL: STEP-SPECIFIC OVER-SEARCH CALIBRATION FOR SEARCH AGENTS VIA SELF-DISTILLED HINDSIGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Search agents are prone to over-search when they fail to accurately assess whether their internal knowledge is sufficient, resulting in unnecessary external search. Outcome-based RL can further amplify this miscalibration, as trajectory-level credit does not distinguish between useful and unnecessary search decisions. Existing process-aware methods incorporate process-level feedback on search necessity, yet the degree of correction applied to individual search decisions remains largely determined by predefined reward formulations or advantage-rescaling rules. This raises a natural question: can the agent itself provide an adaptive, decision-specific signal for calibrating its search behavior? To this end, we introduce **SOCAL** (**S**tep-Specific **O**ver-Search **Cal**ibration for Search Agents via Self-Distilled Hindsight), which transforms over-search diagnoses into step-specific hindsight and leverages self-distillation for adaptive search calibration. Specifically, SOCAL first performs knowledge probing over sampled trajectories to extract hindsight reflecting the model's knowledge sufficiency. For each diagnosed step, it contrasts the policy's support for the same sampled step under the original and hindsight-conditioned views. The resulting likelihood discrepancy serves as a self-distilled signal for reshaping the corresponding outcome-derived credit, allowing the correction strength to adapt to each individual search step. Extensive experiments on seven knowledge-intensive question-answering benchmarks demonstrate the effectiveness of SOCAL, yielding 1.5–2.9 percentage-point gains in CEM over the strongest baselines and the lowest over-search rates across both 3B and 7B model settings. Our code is publicly available at `https://anonymous.4open.science/r/SOCAL`.

1 INTRODUCTION

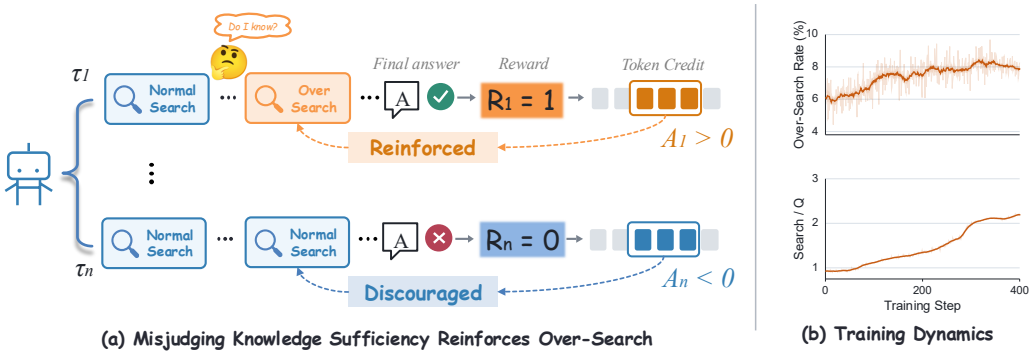

Figure 1: **Trajectory-level RL can reinforce over-search.** (a) A misjudgment of knowledge sufficiency triggers an unnecessary search. Trajectory-level feedback reinforces this search when the final answer is correct. (b) As a representative trajectory-level RL search agent, Search-R1 shows increasing over-search rate and search frequency during training.

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.

Search-augmented language agents interleave internal reasoning with external search, allowing language models to acquire evidence beyond their parametric knowledge (Xi et al., 2026; Li et al., 2025a;b). Recent systems further use reinforcement learning (RL) to learn search behavior directly from task outcomes (Jin et al., 2025; Song et al., 2025b). Yet effective search depends not only on the ability to retrieve useful information, but also on recognizing when retrieval is unnecessary. When an agent fails to accurately assess whether its internal knowledge is sufficient, it may invoke external search unnecessarily, resulting in *over-search* (Xie et al., 2026b; Wu et al., 2026; Zhang et al., 2026b). Such redundant retrieval increases interaction cost and latency, and may further expose the model to irrelevant or misleading evidence.

Outcome-based RL can further amplify this miscalibration because task-level feedback does not distinguish between useful and unnecessary search decisions. For example, Search-R1 (Jin et al., 2025) trains multi-turn search with GRPO (Shao et al., 2024) using final-answer rewards. Search decisions within the same trajectory therefore inherit the same outcome-level credit regardless of their actual contribution to the answer. As illustrated in Figure 1(a), an unnecessary search in a successful trajectory can be reinforced together with decisions that are genuinely useful. Conversely, useful decisions may receive negative credit when the overall trajectory fails. This credit ambiguity can accumulate over training: Figure 1(b) shows that outcome-only optimization increases both search frequency and the over-search rate, even as task performance improves.

To mitigate this trajectory-level credit ambiguity, recent work has introduced process-level feedback on intermediate search behavior. HiPRAG (Wu et al., 2026) evaluates the necessity of individual search steps and aggregates these judgments into a trajectory-level process reward, whereas LeTS (Zhang et al., 2025) uses stepwise process rewards to modulate the outcome advantage at each think-and-search step. These methods demonstrate the value of search-specific supervision, but the strength of correction for each decision remains largely determined by predefined reward formulations or rescaling rules. Consequently, such supervision may not fully reflect how much a particular search decision should be corrected under the policy's current knowledge state. On-policy distillation offers a complementary way to derive such decision-dependent signals from changes in the model's own predictive distribution (Agarwal et al., 2024; Gu et al., 2024; Xu et al., 2025). OPSD (Zhao et al., 2026a) further shows that the same model can provide informative supervision under different conditioning contexts. This leads to a key question: **can such policy-derived signals be localized to individual over-search decisions and used for adaptive search calibration?**

To address this question, we introduce **SOCAL** (**S**tep-Specific **O**ver-Search **Cal**ibration with Self-Distilled Hindsight). The key idea is to use an over-search diagnosis not as the correction itself, but as a cue for constructing step-specific hindsight. For each diagnosed search decision, knowledge probing provides model-specific evidence about knowledge availability, which is distilled into localized hindsight for that step. Rather than assigning a predetermined correction from the diagnosis alone, SOCAL evaluates the same sampled step under the original and hindsight-conditioned views using the same policy. The resulting likelihood change reflects how the added hindsight shifts the policy's support for the original decision, providing a self-distilled signal for determining the strength of local advantage reshaping. In this way, SOCAL adapts credit correction to the characteristics of each individual over-search decision. Our contributions are threefold:

- We introduce SOCAL, which transforms over-search diagnoses into step-specific hindsight that captures the model's knowledge availability, providing a richer representation of each diagnosed search decision than the diagnosis alone.
- We derive adaptive, decision-specific calibration signals from the policy's distributional changes under hindsight conditioning, enabling localized credit correction for individual over-search decisions.
- Extensive experiments on seven question-answering benchmarks show that SOCAL consistently improves answer quality while reducing over-search, indicating that better search calibration need not come at the cost of task performance.

2 Related Work

Search-augmented language agents. Search-augmented language agents extend retrieval-augmented generation from a fixed retrieve-then-answer pipeline to a multi-turn process of plan-

ning, querying, evidence inspection, and reasoning revision (Xi et al., 2026). Early systems established interleaved reasoning and retrieval in fixed corpora and on the open web (Li et al., 2025a;b), while subsequent methods learned search policies from task-level rewards (Jin et al., 2025; Chen et al., 2025; Zheng et al., 2025). Research in 2026 has increasingly focused on scalable training infrastructure and longer search horizons. SearchGym (Zhang et al., 2026c) constructs aligned simulated environments for curriculum RL, whereas OpenSeeker (Du et al., 2026) builds fact-grounded tasks and denoised trajectories for open-source SFT. SearchArt (Mei et al., 2026) couples verified task synthesis with multi-stage post-training for deep-search and deep-research settings. In parallel, CaRR (Zhang et al., 2026a) augments binary success signals with citation-aware rubric rewards that promote complete evidence chains. GrepSeek (Salemi et al., 2026) broadens the interaction interface by training agents to search corpora through executable commands rather than a fixed retriever. Together, these developments shift the field from eliciting basic tool use toward training search agents over controllable environments, richer feedback, and extended interaction horizons.

On-policy distillation for language agents. On-policy distillation (OPD) matches a teacher distribution on student-generated trajectories, providing token-level supervision on states encountered by the student (Agarwal et al., 2024; Gu et al., 2024; Xu et al., 2025). On-policy self-distillation (OPSD) later uses the same model as a privileged teacher and an unprivileged student (Zhao et al., 2026a). Recent variants adapt this framework to long-horizon agents by selecting or restructuring the Teacher signal. AOPD (Jia et al., 2026) and ROSD (Zhao et al., 2026b) condition distillation on policy advantage or reflected errors. SDAR (Lu et al., 2026), TurnOPD (Zhou et al., 2026), and ICSD (Lan et al., 2026) refine its allocation across tokens and turns. TRACE (Wang et al., 2026a) routes KL to annotated error spans, leaves other tokens under GRPO, and later closes the privileged channel. AgentOPSD (Wang et al., 2026b) instead recursively aggregates token-level discrepancies into turn-level evidence. Search-E1 (Liang et al., 2026), KbSD (Feng et al., 2026), and LoongReflect (Zhang et al., 2026d) apply self-distillation to search through sibling trajectories, knowledge-boundary context, and reflection traces. EviSD (Xie et al., 2026a) uses supporting evidence and gold answers as privileged context, converting token-level likelihood differences into bounded adjustments of action-token advantages.

Adaptive retrieval and search-decision calibration. Early work on retrieval control decides when additional evidence is needed at inference time. FLARE (Jiang et al., 2023) triggers retrieval from low-confidence predictions, while Self-RAG (Asai et al., 2024) learns reflection tokens that control retrieval and evidence use. Adaptive-RAG (Jeong et al., 2024) routes queries among retrieval strategies, and DRAGIN (Su et al., 2024) estimates information needs during generation. More recent work directly trains search decisions. LeTS (Zhang et al., 2025) adds process rewards to outcome RL, and HiPRAG (Wu et al., 2026) aggregates necessity judgments for search and non-search steps. Other work characterizes over-search (Xie et al., 2026b), while DAS (Zhang et al., 2026b) and BOUND (Niu et al., 2026) construct local preferences for decision calibration. Together, these studies move retrieval control from inference-time triggers toward learned supervision for individual search decisions. Within the latter line, supervision takes the form of process rewards, necessity judgments, or local preferences.

3 PRELIMINARIES

We consider a search-augmented agent that performs multi-step reasoning with access to an external search environment. Given a question x, the agent generates a trajectory

$$\boldsymbol{\tau} = (x, s_1, \ldots, s_K, y), \tag{1}$$

where K is the number of interaction steps, s_k denotes the k-th step, and y is the final answer.

Following the step-structured interaction adopted in prior search-agent training (Wu et al., 2026), each step is either a search step or a non-search step. A search step consists of policy-generated reasoning, a search query q_k, the returned context, and a policy-generated conclusion z_k, whereas a non-search step contains reasoning and a conclusion without invoking external search.

We denote the set of search-step indices as

$$\mathcal{S} = \{k \in \{1, \ldots, K\} : s_k \text{ invokes search}\}, \tag{2}$$

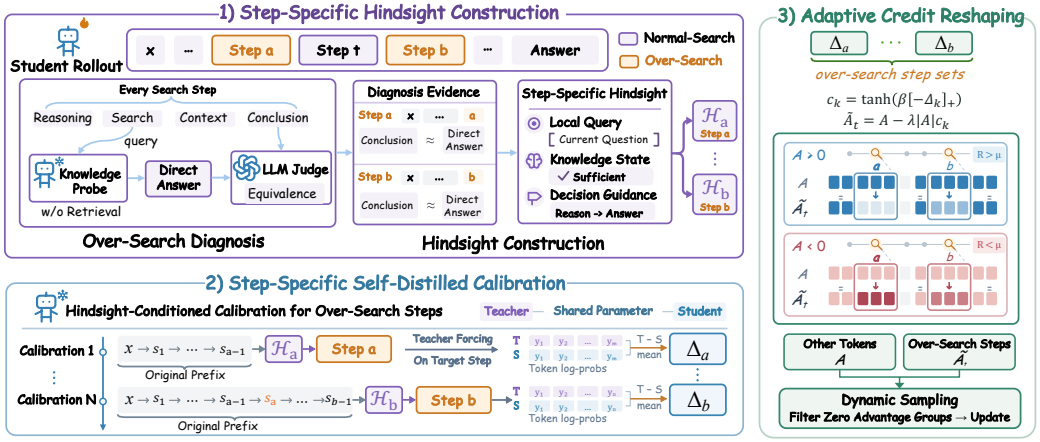

Figure 2: **Overview of SOCAL.** (1) **Step-Specific Hindsight Construction:** knowledge probing identifies over-search decisions and extracts localized hindsight reflecting the model's knowledge Sufficiency. (2) **Step-Specific Self-Distilled Calibration:** each diagnosed over-search step is independently evaluated under its original and hindsight-conditioned views, yielding a step-specific calibration signal. (3) **Adaptive Credit Reshaping:** the self-distilled signal determines the correction strength for each diagnosed over-search decision and locally reshapes its credit.

and define the trajectory prefix preceding step k as

$$h_k = [x; s_1; \ldots; s_{k-1}]. \tag{3}$$

4 METHOD

In this section, we present SOCAL, a step-specific framework for calibrating over-search in search-agent RL. As illustrated in Figure 2, SOCAL proceeds in three stages: Step-Specific Hindsight Construction (Section 4.1), Step-Specific Self-Distilled Calibration (Section 4.2), and Adaptive Credit Reshaping (Section 4.3). Together, these stages derive a localized calibration signal for each diagnosed over-search decision, enabling adaptive adjustment of its assigned credit.

4.1 STEP-SPECIFIC HINDSIGHT CONSTRUCTION

Over-search diagnosis. A sampled search step alone does not reveal whether external search was necessary for obtaining its post-search conclusion. Following HiPRAG (Wu et al., 2026), we therefore perform knowledge probing to identify potentially unnecessary searches. For each search step $k \in \mathcal{S}$, we present the corresponding query q_k to the current policy without external search, obtaining a direct response $\hat{z}_k$. An LLM judge then compares $\hat{z}_k$ with the original post-search conclusion z_k for semantic equivalence. We define the diagnosed over-search set as

$$\mathcal{O} = \{k \in \mathcal{S} : \mathcal{J}_{\text{LLM}}(\hat{z}_k, z_k) = 1\}, \tag{4}$$

where $\mathcal{J}_{\text{LLM}}$ denotes the semantic-equivalence judgment. Agreement indicates that the search may have been unnecessary, serving as evidence of knowledge Sufficiency rather than an independent guarantee of answer correctness.

Step-specific hindsight. For each diagnosed step $k \in \mathcal{O}$, we construct a step-specific hindsight $\mathcal{H}_k$ that summarizes the evidence relevant to the corresponding search decision:

- **Local query:** specifies the local information need that the current search decision is intended to resolve, making explicit what information the agent is seeking at this step.
- **Knowledge state:** records whether the model's existing knowledge appears sufficient for the current information need, together with the supporting evidence inferred from the semantic agreement between the no-search response and the post-search conclusion.

- **Decision guidance:** converts the sufficiency assessment into localized guidance for the current search decision, encouraging the agent to reason toward an answer rather than invoke external search when the required knowledge appears sufficient.

Each diagnosed step is associated with its own hindsight $\mathcal{H}_k$. To keep the hindsight focused on the search decision itself, we exclude the probe response, the original post-search conclusion, and the retrieved context. Thus, $\mathcal{H}_k$ retains only the local information need, evidence of knowledge sufficiency, and the corresponding decision guidance, which will serve as the conditioning information for the subsequent calibration stage.

4.2 Step-Specific Self-Distilled Calibration

For each diagnosed over-search step, SOCAL derives a separate calibration signal from its corresponding hindsight. A trajectory with $N = |\mathcal{O}|$ diagnosed steps therefore yields N independent calibration instances, each evaluating the same sampled step under two views. Specifically, for step $k \in \mathcal{O}$, the *original view* (Student view) uses the original trajectory context, whereas the *hindsight-conditioned view* (Teacher view) additionally conditions on the corresponding hindsight $\mathcal{H}_k$:

$$v_k^{\mathrm{O}} = [h_k; s_k], \qquad v_k^{\mathrm{H}} = [h_k; \mathcal{H}_k; s_k]. \tag{5}$$

Both views are evaluated by the same pre-update policy $\pi_{\theta_{\mathrm{old}}}$ and differ only in hindsight conditioning. Following the self-distillation principle of privileged-context scoring (Zhao et al., 2026a), their distributional difference provides localized feedback for the over-search decision being calibrated.

We next compute this difference at the token level. For the t-th policy-generated token $y_{k,t}$ in step k, let $s_{k,<t}$ denote the preceding content within the same step. The two views assign

$$\begin{aligned} p_{k,t}^{\mathrm{O}} &= \pi_{\theta_{\mathrm{old}}}(y_{k,t} \mid [h_k; s_{k,<t}]), \\ p_{k,t}^{\mathrm{H}} &= \pi_{\theta_{\mathrm{old}}}(y_{k,t} \mid [h_k; \mathcal{H}_k; s_{k,<t}]). \end{aligned} \tag{6}$$

If a returned search context appears before $y_{k,t}$, it remains part of the conditioning sequence, but is excluded from the policy-generated tokens used to construct the calibration signal.

Step-specific calibration signal. Let $\mathcal{M}_k$ denote the policy-generated token positions in step k, including its reasoning, query, and conclusion tokens while excluding the returned search context. We aggregate the hindsight-induced likelihood change over these tokens as

$$\Delta_k = \mathrm{sg}\left[\frac{1}{|\mathcal{M}_k|}\sum_{t\in\mathcal{M}_k}\left(\log p_{k,t}^{\mathrm{H}} - \log p_{k,t}^{\mathrm{O}}\right)\right], \tag{7}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Averaging within the step reduces dependence on step length, preventing longer steps from receiving larger-magnitude signals simply because they contain more generated tokens.

The resulting Δ_k measures how hindsight conditioning changes the policy's support for the sampled step. A negative Δ_k indicates that the hindsight-conditioned view assigns lower likelihood to the original step, with a larger negative gap reflecting stronger disagreement with the sampled decision. We therefore use the negative part of Δ_k as a self-distilled signal to determine the correction strength in the subsequent credit-reshaping stage.

4.3 Adaptive Credit Reshaping

Given a group of G trajectories sampled for the same question, GRPO computes a group-relative outcome advantage for each trajectory i (Shao et al., 2024):

$$A_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon}, \tag{8}$$

where μ_R and σ_R denote the mean and standard deviation of rewards within the group. Because the same A_i is shared across all policy-generated tokens in a trajectory, a diagnosed over-search step may receive positive credit simply because the final outcome is successful. We therefore use the

step-specific calibration signal $\Delta_{i,k}$ to adaptively reshape the credit assigned to each diagnosed over-search step. Specifically, we convert the negative part of $\Delta_{i,k}$ into a bounded correction strength:

$$c_{i,k} = \tanh(\beta[-\Delta_{i,k}]_+)\,, \tag{9}$$

where $[\cdot]_+$ activates the correction only when the hindsight-conditioned view assigns lower support to the sampled step. The $\tanh$ mapping allows stronger disagreement to induce a larger correction while keeping its magnitude bounded. We then reshape the token-level advantage as

$$\widetilde{A}_{i,t} = \begin{cases} A_i - \lambda |A_i| c_{i,k}, & k \in \mathcal{O}_i,\ t \in \mathcal{M}_{i,k}, \\ A_i, & \text{otherwise.} \end{cases} \tag{10}$$

Scaling the correction by $|A_i|$ keeps it proportional to the original outcome signal. Consequently, the reshaping attenuates positive credit and strengthens negative credit for diagnosed over-search steps, while leaving the remaining decisions unchanged. The coefficient λ controls the maximum relative correction.

Dynamic sampling and policy optimization. Because the correction in Eq. 10 scales with $|A_i|$, it vanishes for uniform-reward groups whose group-relative advantages are zero. We therefore use dynamic sampling to reduce such groups and increase the number of effective credit-reshaping updates. Finally, we optimize the policy with the standard clipped GRPO objective using $\widetilde{A}_{i,t}$:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\min\Big(r_{i,t}(\theta)\widetilde{A}_{i,t}, \mathrm{clip}(r_{i,t}(\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})\,\widetilde{A}_{i,t}\Big)\,, \tag{11}$$

where $\mathcal{T}_i$ denotes the policy-generated tokens in trajectory i, and

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid \boldsymbol{\tau}_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid \boldsymbol{\tau}_{i,<t})}. \tag{12}$$

5 Experiments

In this section, we evaluate SOCAL from three complementary perspectives: answer correctness, search calibration, and component effectiveness. We compare it with existing search-augmented methods, examine its ability to reduce unnecessary search without compromising useful retrieval, and analyze the contribution of its key design choices.

5.1 Experimental Setup

Datasets. We train SOCAL using the official training splits of NQ (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018), covering both factual and multi-hop question answering during training. Evaluation is conducted on all seven benchmarks, including the corresponding evaluation splits of NQ and HotpotQA as well as five additional datasets that are never seen during training. This setup allows us to assess both in-distribution performance and cross-dataset generalization. In total, the evaluation set contains 51,713 examples.

Metrics. We evaluate both answer correctness and search behavior. For answer quality, we use Cover Exact Match (CEM). We extract the final text enclosed by `<answer>` tags, apply standard normalization, and count a prediction as correct if it contains any normalized gold answer. We report per-dataset CEM and the average CEM over all evaluation examples.

For search behavior, we adopt the Over-search Rate (OSR) from HiPRAG (Wu et al., 2026). Let S_i denote the number of search calls for question i and O_i the number judged unnecessary. Over N valid trajectories,

$$\mathrm{OSR} = \frac{\sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} S_i}, \qquad \mathrm{Search/Q} = \frac{1}{N}\sum_{i=1}^{N} S_i. \tag{13}$$

OSR measures the proportion of unnecessary search decisions, while Search/Q reflects overall search frequency. Reporting both helps distinguish better search calibration from simply reducing the number of searches.

Method	Single-hop			Multi-hop				Avg. CEM
	NQ	TriviaQA	PopQA	HotpotQA	2WikiMQA	MuSiQue	Bamboogle	
Inference- and prompt-based methods								
Direct Inference	27.0	26.8	40.1	58.7	16.0	7.9	15.9	31.8
Standard RAG (Lewis et al., 2020)	51.2	54.7	65.7	56.9	21.6	18.5	18.6	45.3
IRCoT (Trivedi et al., 2023)	27.5	36.0	42.5	51.4	37.6	19.4	20.6	36.4
Search-o1 (Li et al., 2025a)	40.2	42.2	58.5	56.1	45.6	15.1	19.3	43.9
RL-based search agents								
R1-Searcher (Song et al., 2025a)	60.0	73.0	58.2	60.4	60.3	32.9	55.8	60.6
R1-Searcher++ (Song et al., 2025b)	61.0	73.5	59.0	64.2	63.2	32.3	**58.7**	62.1
β-GRPO (Wu et al., 2025)	65.0	75.0	60.0	53.0	66.0	24.0	52.0	62.5
β-GRPO*	62.4	73.9	61.3	52.1	66.0	22.8	54.4	62.1
Search-R1-3B (Jin et al., 2025)*	64.8	73.8	60.6	55.4	66.0	26.5	37.6	62.7
Search-R1-7B*	69.3	76.8	62.1	61.9	67.1	33.7	44.8	65.7
Process-aware RL search agents								
HiPRAG-3B (Wu et al., 2026)*	68.7	75.5	**66.3**	57.4	67.4	24.1	41.6	65.4
HiPRAG-7B*	71.2	76.3	63.2	62.4	71.7	34.1	52.8	67.2
SOCAL-3B	69.4	77.1	63.0	62.7	70.1	35.4	51.2	66.9
SOCAL-7B	**73.4**	**79.2**	65.2	**67.7**	**74.0**	**38.6**	54.4	**70.1**

Table 1: **Main results** (CEM, %) on seven single-hop and multi-hop QA benchmarks. * denotes methods evaluated with the same prompt template as SOCAL; unmarked methods use their original default prompts. The best result in each column is shown in **bold**. For clarity, we omit the * notation in the following discussion and refer to the prompt-aligned variants simply by their method names.

Baselines. We compare SOCAL with a broad range of search-augmented methods. **Direct Inference** answers questions without external search, while **Standard RAG** (Lewis et al., 2020) retrieves evidence before generation. **IRCoT** (Trivedi et al., 2023) and **Search-o1** (Li et al., 2025a) represent prompt-based approaches that interleave reasoning with search. For RL-based search agents, we include **R1-Searcher** (Song et al., 2025a), **R1-Searcher++** (Song et al., 2025b), **Search-R1** (Jin et al., 2025), and **β-GRPO** (Wu et al., 2025). We further compare with **HiPRAG** (Wu et al., 2026), which incorporates step-level search-necessity judgments through process rewards during RL training.

Implementation Details. We train Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct with GRPO using veRL. The rollout group size is set to $G = 5$, with a global batch size of 512 prompts per optimization step. For retrieval, we use E5-base-v2 over the 2018 Wikipedia corpus, returning the top-3 passages for each search query. The search budget is capped at four calls per trajectory. Following HiPRAG, baselines trained in our pipeline use a 400-step training budget. We train SOCAL for 320 steps because its training metrics have largely converged by then, while later dynamic sampling becomes increasingly expensive for obtaining groups with nonzero advantages. For SOCAL, we set the credit-reshaping coefficient to $\lambda = 0.2$ and the likelihood-gap scale to $\beta = 5$. For search-behavior evaluation, trajectories are generated by the checkpoint being evaluated, and each search decision is assessed by a fixed `qwen3.7-flash` evaluator for retrieval necessity. The evaluator and its prompt remain unchanged throughout all experiments.

5.2 Main Results

Table 1 reports answer correctness across seven knowledge-intensive QA benchmarks. From the results, we make the following observations.

Strong performance across model scales. SOCAL consistently improves answer quality at both model scales. In the 3B setting, SOCAL achieves an average CEM of 66.9, outperforming Search-R1 and HiPRAG by 4.2 and 1.5 points, respectively. In the 7B setting, the average CEM further increases to 70.1, exceeding the corresponding baselines by 4.4 and 2.9 points. These results show that the effectiveness of SOCAL remains consistent as model capacity increases.

Consistent gains across tasks and unseen benchmarks. The improvement is not confined to a particular dataset or question type. At the 7B scale, SOCAL outperforms HiPRAG on all seven benchmarks and achieves the best result on five of them. The gains span both single-hop and multi-

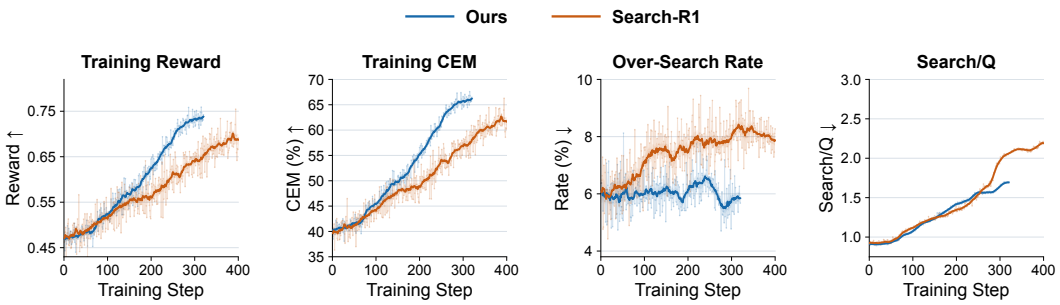

Figure 3: **Training dynamics of SOCAL and Search-R1 (3B).** We show the evolution of training reward, answer correctness, over-search rate, and search frequency.

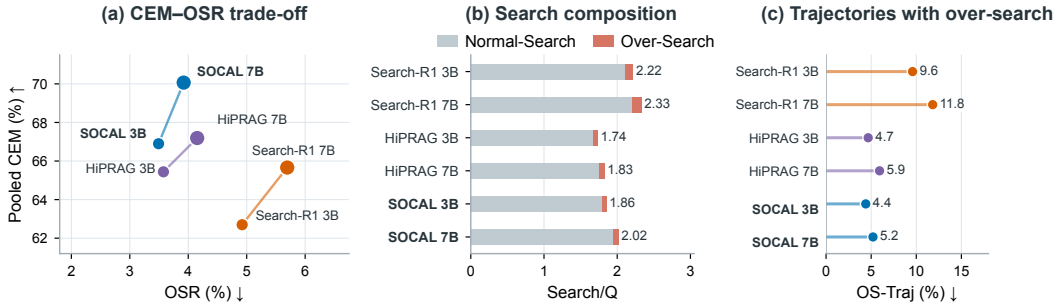

Figure 4: **Search decision analysis across methods.** Left: average CEM vs OSR. Middle: Search/Q decomposed into over-searches and normal-earches. Right: the percentage of trajectories containing at least one unnecessary search (OS-Traj).

hop questions and extend to the five datasets unseen during training, suggesting that the improvement generalizes beyond the training benchmarks.

5.3 Search Decision Analysis

We further examine how SOCAL affects search decisions during training and at evaluation. Figure 3 compares the training dynamics of SOCAL and Search-R1, while Figure 4 evaluates their final search behavior together with HiPRAG. We additionally report OS-Traj, the percentage of valid trajectories containing at least one unnecessary search.

SOCAL mitigates the accumulation of over-search during training. As shown in Figure 3, both methods improve in reward and answer correctness as training proceeds, but their search behavior evolves differently. Search-R1 exhibits an increase in both OSR and Search/Q, suggesting that outcome-based optimization may reinforce unnecessary searches together with successful trajectories. In contrast, SOCAL achieves stronger improvements in reward and CEM while keeping OSR more stable and search frequency lower. This trend supports the effectiveness of step-specific calibration in preventing over-search from being progressively amplified during training.

SOCAL improves search calibration rather than simply reducing search. Figure 4 shows that the reduction in over-search persists at evaluation. Compared with Search-R1, SOCAL reduces OSR from 4.92% to 3.49% at 3B and from 5.69% to 3.92% at 7B. The same pattern is more pronounced at the trajectory level: OS-Traj decreases from 9.59% to 4.40% and from 11.82% to 5.20%, respectively. SOCAL also achieves lower OSR and OS-Traj than HiPRAG at both model scales.

Importantly, this improvement does not come from simply minimizing search frequency. SOCAL performs more searches than HiPRAG at both model scales, while still achieving lower OSR and higher CEM. The decomposition of Search/Q further shows that nearly all of these additional

Variant	Hindsight scope	Placement	Credit	CEM ↑	OSR ↓	Search/Q
Hindsight construction and conditioning						
w/ trajectory-prefix hindsight	Step	Prefix	Step	65.35	4.86	2.06
w/ global hindsight	Global	Prefix	Step	64.75	4.63	1.97
w/ action-only hindsight	Step	Target	Step	63.95	5.24	1.70
Credit adjustment and sampling						
w/ token-level adjustment	Step	Target	Token	66.62	3.66	1.91
w/o dynamic sampling	Step	Target	Step	65.11	4.38	2.02
Full SOCAL	**Step**	**Target**	**Step**	**66.90**	**3.49**	1.86

Table 2: **Ablation studies of SOCAL (3B).** "Hindsight scope" distinguishes *Step*, which assigns separate hindsight to each diagnosed over-search step, from *Global*, which shares hindsight across steps. "Placement" indicates whether hindsight is placed at the trajectory *Prefix* or immediately before the target step in the Teacher view. "Credit" denotes the granularity of advantage adjustment.

searches are judged necessary. Together, these results indicate that effective search optimization is not about searching less, but about better calibrating when external information is actually needed.

5.4 Ablation Studies

We conduct ablation studies on the 3B model to examine three key design choices of SOCAL: the information contained in self-distilled hindsight, how hindsight is localized to individual over-search steps, and how the resulting signal is used for policy optimization.

Informative hindsight matters. We first replace the full hindsight with action-only guidance, removing query-specific and knowledge-availability evidence. Although this variant achieves the lowest Search/Q (1.70), its CEM and OSR degrade substantially to 63.95 and 5.24%, respectively. This suggests that simply encouraging the policy to search less is insufficient; effective calibration requires hindsight that captures both the current information need and the evidence that external search may be unnecessary.

Hindsight should be localized and step-specific. We next vary both the placement and scope of hindsight. Moving step-specific hindsight to the trajectory prefix reduces CEM to 65.35 and increases OSR to 4.86%, while replacing independent step-level hindsight with a shared global context further lowers CEM to 64.75. These results support associating each diagnosed over-search step with its own hindsight and conditioning it locally at the corresponding target step.

Step-level credit adjustment and dynamic sampling improve optimization. Applying the correction at the token level remains competitive but is slightly weaker than the step-level design in both CEM and OSR. Removing dynamic sampling causes a clearer degradation, reducing CEM to 65.11 and raising OSR to 4.38%. This agrees with the motivation in Section 4.3: reducing uniform-reward groups increases the number of effective credit-reshaping updates, which is important for ensuring that the hindsight signal can meaningfully influence policy optimization.

6 Conclusion

We introduce SOCAL, a step-specific framework for mitigating over-search in search-augmented language agents through self-distilled hindsight. SOCAL transforms over-search diagnoses into localized hindsight and uses the resulting distributional feedback to adaptively reshape the credit assigned to individual search decisions. Experiments on seven knowledge-intensive QA benchmarks demonstrate consistent improvements in answer quality together with lower over-search across model scales. Further analysis shows that these gains come from better search calibration rather than indiscriminate retrieval suppression, while ablations confirm the importance of informative, step-specific hindsight and localized credit reshaping. Overall, our results suggest that effective search-agent training should account not only for trajectory-level outcomes, but also for whether individual search decisions are actually necessary.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, figure ideation and generation, and research-code development. The authors reviewed all AI-assisted text and visual artifacts, verified cited literature against primary sources, and inspected and tested AI-assisted code before use. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. In *Advances in Neural Information Processing Systems 38*, pp. 94974–94994. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-2858. URL https://doi.org/10.52202/085713-2858.

Yuwen Du, Rui Ye, Shuo Tang, Xinyu Zhu, Yijun Lu, Yuzhu Cai, and Siheng Chen. OpenSeeker: Democratizing frontier search agents by fully open-sourcing training data, 2026. URL https://arxiv.org/abs/2603.15594.

Tao Feng, Xinke Jiang, and Chao Wu. KbSD: Knowledge boundary aware self-distillation for behavioral calibration in agentic search, 2026. URL https://arxiv.org/abs/2606.29863.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level, 2026. URL https://arxiv.org/abs/2605.06387.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

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. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Qizhen Lan, Xi Xiao, Xiangchen Guan, Mengchen Fan, Moule Lin, Jung Im Choi, and Lijing Zhu. Trust is not enough: Influence calibration for on-policy self-distillation in agentic RL, 2026. URL https://arxiv.org/abs/2608.14945.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen tau Yih, Tim Rockt"aschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. 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, 2025b. URL https://arxiv.org/abs/2504.21776.

Zihan Liang, Yufei Ma, Ben Chen, Zhipeng Qian, Xuxin Zhang, Huangyu Dai, and Lingtao Mao. Search-E1: Self-distillation drives self-evolution in search-augmented reasoning, 2026. URL https://arxiv.org/abs/2605.22511.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-distilled agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2605.15155.

Lang Mei, Xiaohan Yu, Chong Chen, Liyan Liu, Xiangnan Chen, Jinchao Ma, Chao Feng, Li Huang, Siyu Mo, Sichen Kang, Yunkun Xu, Zhihan Yang, Zhujun Xue, Jingren Zhang, Qing He, Yingdi Huang, Hao Jiang, Ziao Ma, Zewei Pan, Minhao Sun, Zhuo Tao, Jinzhao Xiao, Gangtao Xin, Huanyao Zhang, Wenjian Zhang, Jiangshan Zhang, Guojie Zhu, Fangzhou Zou, Jiaxin Mao, and Wentao Zhang. SearchArt: Training long-horizon search agent with scalable synthetic and verified task, 2026. URL https://arxiv.org/abs/2607.24850.

Qingying Niu, Ruiyang Ren, Wayne Xin Zhao, and Yaliang Li. BOUND: Brief-guided corrective preference distillation at search-control boundaries, 2026. URL https://arxiv.org/abs/2608.08768.

Alireza Salemi, Chang Zeng, Atharva Nijasure, Jui-Hui Chung, Razieh Rahimi, Fernando Diaz, and Hamed Zamani. GrepSeek: Training search agents for direct corpus interaction, 2026. URL https://arxiv.org/abs/2605.29307.

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

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning, 2025a. URL https://arxiv.org/abs/2503.05592.

Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning, 2025b. URL https://arxiv.org/abs/2505.17005.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12991–13013, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.702. URL https://aclanthology.org/2024.acl-long.702/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Jiaxuan Wang, Xuan Ouyang, Zhiyu Chen, Yulan Hu, Zheng Pan, Xin Li, and Lan-Zhe Guo. TRACE: Distilling where it matters via token-routed self on-policy alignment, 2026a. URL https://arxiv.org/abs/2605.10194.

Zi-Han Wang, Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Jie Wu, Zhengzhou Cai, Yueqing Sun, Ziang Ye, Linji Hao, Qi Gu, Xunliang Cai, Yongliang Shen, and Yujiu Yang. AgentOPSD: Recursive self-distillation for agentic reinforcement learning, 2026b. URL https://arxiv.org/abs/2608.05987.

Peilin Wu, Mian Zhang, Xinlu Zhang, Xinya Du, and Zhiyu Zoey Chen. Search wisely: Mitigating sub-optimal agentic searches by reducing uncertainty. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19723–19734, 2025. doi: 10.18653/v1/2025.emnlp-main.998. URL https://aclanthology.org/2025.emnlp-main.998/.

Peilin Wu, Mian Zhang, Kun Wan, Wentian Zhao, Kaiyu He, Xinya Du, and Zhiyu Chen. Hiprag: Hierarchical process rewards for efficient agentic retrieval augmented generation, 2026. URL https://arxiv.org/abs/2510.07794.

Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu, and Weinan Zhang. A survey of large language model-based search agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8244–8279, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.374. URL https://aclanthology.org/2026.acl-long.374/.

Jianan Xie, Xin Sun, Zhongqi Chen, Xing Zheng, Shu Wu, Bowen Song, and Liang Wang. EviSD: Evidence-conditioned self-distillation for search-augmented agents, 2026a. URL https://arxiv.org/abs/2608.01359.

Roy Xie, Deepak Gopinath, David Qiu, Dong Lin, Haitian Sun, Saloni Potdar, and Bhuwan Dhingra. Over-searching in search-augmented large language models, 2026b. URL https://arxiv.org/abs/2601.05503.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=EgJhwYR2tB.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Jiajie Zhang, Xin Lv, Ling Feng, Lei Hou, and Juanzi Li. Chaining the evidence: Robust reinforcement learning for deep search agents with citation-aware rubric rewards, 2026a. URL https://arxiv.org/abs/2601.06021.

Qi Zhang, Shouqing Yang, Lirong Gao, Hao Chen, Xiaomeng Hu, Jinglei Chen, Jiexiang Wang, Sheng Guo, Bo Zheng, Haobo Wang, and Junbo Zhao. Lets: Learning to think-and-search via process-and-outcome reward hybridization, 2025. URL https://arxiv.org/abs/2505.17447.

Wenlin Zhang, Kuicai Dong, Junyi Li, Yingyi Zhang, Xiaopeng Li, Pengyue Jia, Yi Wen, Derong Xu, Maolin Wang, Yichao Wang, Yong Liu, and Xiangyu Zhao. To search or not to search: Aligning the decision boundary of deep search agents via causal intervention, 2026b. URL https://arxiv.org/abs/2602.03304.

Xichen Zhang, Ziyi He, Yinghao Zhu, Sitong Wu, Shaozuo Yu, Meng Chu, Wenhu Zhang, Haoru Tan, and Jiaya Jia. SearchGym: Bootstrapping real-world search agents via cost-effective and high-fidelity environment simulation, 2026c. URL https://arxiv.org/abs/2601.14615.

Zhixin Zhang, Xinke Jiang, Zhibang Yang, Weixuan Xu, Guohong Qiu, Xu Chu, Junfeng Zhao, and Yasha Wang. LoongReflect: Boosting long-horizon reflection in search agents via global perspective distillation, 2026d. URL https://arxiv.org/abs/2608.11967.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026a. URL https://arxiv.org/abs/2601.18734.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. ROSD: Reflective on-policy self-distillation for language model reasoning across domains, 2026b. URL https://arxiv.org/abs/2605.28014.

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, 2025. URL https://arxiv.org/abs/2504.03160.

Yuhang Zhou, Kai Zheng, Haoling Li, Dengyun Peng, Can Xu, and Jingjing Chen. TurnOPD: Making on-policy distillation turn-aware for efficient long-horizon agent training, 2026. URL https://arxiv.org/abs/2607.05804.