CAFE: SELF-IMPROVING SEARCH AGENTS NEED CO-EVOLVING FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable search requires more than acquiring external evidence. An agent must also recognize and recover from errors as its trajectory unfolds. In-trajectory feedback provides a mechanism for such recovery by diagnosing where the search has drifted and redirecting subsequent reasoning steps. This is particularly important in long-horizon search, where an early directional error may receive no immediate corrective signal and can compound across later steps. Making such feedback learnable, however, creates a coupled problem: the agent must learn when to request and use feedback, while the critic must learn corrections from outcome-confounded rollouts as the agent's failure patterns evolve. We introduce CAFE (Coupled Agent–Feedback Evolution), a framework in which a shared-parameter model alternates between search-agent and critic roles. CAFE initializes feedback-conditioned recovery from trajectories built around the base agent's own failures, then couples online and offline optimization. During online RL, a comparative feedback estimate uses a prompt-level call–skip success gap to shape request returns, while feedback-aware advantage shaping reweights token advantages before and after feedback. Offline, rollout-derived preference optimization learns feedback from matched successful and unsuccessful trajectories. On seven agentic search benchmarks, CAFE outperforms the evaluated RL-based search agents on average, retains its gains across all six out-of-domain benchmarks, and reduces answer-level hallucinations. One-sided ablations show that improving only the agent or only the critic eventually plateaus, whereas alternating the two updates continues to improve performance. These findings suggest that a self-improving search agent needs feedback that co-evolves with the policy it guides.

1 INTRODUCTION

Search agents answer knowledge-intensive questions by iteratively interacting with external search environments, issuing queries and revising their behavior based on retrieved evidence (Xi et al., 2025; Li et al., 2025b; Zheng et al., 2025b). Many earlier retrieval-augmented pipelines relied on fixed strategies that determined when retrieval should occur, either at predetermined reasoning stages (Trivedi et al., 2023) or when a hand-crafted condition was triggered (Jiang et al., 2023). In contrast, outcome-supervised search agents learn policies for when and how to search from terminal task rewards (Jin et al., 2025a; Song et al., 2025). Yet this autonomy remains largely outward-facing, with agents learning what external knowledge to seek while lacking comparable introspection into their own search trajectories.

In long-horizon search, an early directional error may receive no immediate corrective signal. Its cost is not confined to the errant step but propagates throughout the subsequent trajectory (Wang et al., 2026a; An et al., 2026; Qi et al., 2026). Nor is the source of failure easy to identify retrospectively. For instance, the agent may apply a constraint to the initial candidates but silently drop it in later steps, leave a required subgoal unexplored, or repeatedly rephrase a query without acquiring new evidence (Wong et al., 2026; Polshkov et al., 2026). In such scenarios, the trajectory begins correctly but ends in failure, and the terminal reward cannot say where. Recent work has therefore introduced finer-grained credit through information gain (Wang et al., 2025), confidence change (Xie et al., 2026), and local state comparison (Zheng et al., 2025a; Feng et al., 2025). These signals remain evaluative rather than instructive (Sutton et al., 1998), assigning credit retrospectively while deferring correction to future policy updates. Redirecting the active trajectory instead requires in-context feedback that diagnoses where the search has drifted and prescribes what to try next.

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.

Although prior work has explored in-context feedback, methods such as Reflexion (Shinn et al., 2023), Self-Refine (Madaan et al., 2023), and CRITIC (Gou et al., 2024) typically return natural-language critiques post hoc and via prompted rather than trained models. Making such feedback learnable and delivering it within the active trajectory introduces three intertwined challenges. The first lies with the agent, which must learn when to request an intervention, while its learning objective must determine which behavior the resulting success should reinforce, given that a rescued trajectory and a flawless one yield identical rewards (Uesato et al., 2022). The second lies with the critic, whose feedback must be learned without ground truth, from a reward confounded by both the preceding search and the agent's subsequent actions. The third emerges between them, as online updates shift the states the agent visits and the failures it encounters (Ackermann et al., 2025), potentially leaving a static critic misaligned with the evolving policy it is meant to guide.

To address these challenges, we introduce **CAFE** (**C**oupled **A**gent–**F**eedback **E**volution), an iterative, shared-parameter framework that integrates corrective feedback into the active search trajectory and alternates optimization of the agent and critic capabilities. Learning this coupled interaction directly from sparse outcomes is difficult, so we initialize CAFE with recovery demonstrations constructed from the base agent's own failures. Rather than replacing a failed rollout with an ideal trajectory, we preserve its erroneous prefix, insert corrective feedback, and retain a successful continuation. This teaches the model to request, generate, and use feedback at states its own policy actually visits. Imitation alone, however, does not reveal when an intervention is useful or how the resulting success should be credited. During online RL, the comparative feedback estimate (CFE) uses the empirical prompt-level success gap between rollouts with and without feedback requests to shape request returns, while feedback-aware advantage shaping redistributes credit across the intervention. Offline, rollout-derived preference optimization (RDPO) learns from prefix-matched successful and failed trajectories, reducing outcome confounding. At each iteration, CAFE alternates online agent learning with offline critic refinement using the latest rollouts, maintaining alignment between the two capabilities as they co-evolve.

We evaluate CAFE using Qwen2.5-7/3B-Instruct (Qwen et al., 2025) across seven agentic SearchQA benchmarks, with additional evaluation on BrowseComp-Plus (Chen et al., 2025). At the 7B scale, CAFE achieves the strongest average performance among competing RL-based search methods, outperforming the strongest baseline by 2.1 EM and 1.3 F1. Notably, these gains are consistent across all six out-of-domain benchmarks and generalize robustly to the 3B scale. Extensive component and objective ablations confirm the necessity of our feedback-aware credit assignment and the alternating optimization schedule. Furthermore, in-depth analyses of agent–critic cross-play, training dynamics, and hallucination (reducing the average answer-level rate from 17.6% to 12.6%) explicitly demonstrate how sustained co-evolution actively drives the observed performance improvements.

Our contributions are fourfold:

❶ **We formulate self-improving search as a coupled agent–feedback learning problem.** CAFE integrates both roles within a shared-parameter model, alternating online policy updates with offline critic refinement.

❷ **We develop a targeted RL framework for seeking and utilizing feedback.** Our comparative feedback estimate (CFE) and advantage shaping actively learn when to trigger interventions and how to assign credit.

❸ **We introduce rollout-derived preference optimization (RDPO) for feedback generation.** By learning from prefix-matched preference pairs mined from the latest online rollouts, RDPO keeps the critic aligned with the agent as it evolves.

❹ **We provide empirical evidence for co-evolution.** CAFE achieves the best average performance among the evaluated RL-based baselines, transfers robustly to all out-of-domain datasets, reduces hallucinations, and **drives sustained gains in both trajectory quality and search performance**.

2 METHODOLOGY

Motivation. While recent work has primarily advanced search agents' ability to acquire external evidence, we argue that robust long-horizon search also requires active in-trajectory error diagnosis and correction, and as policy updates shift the agent's state and failure distributions, the critic

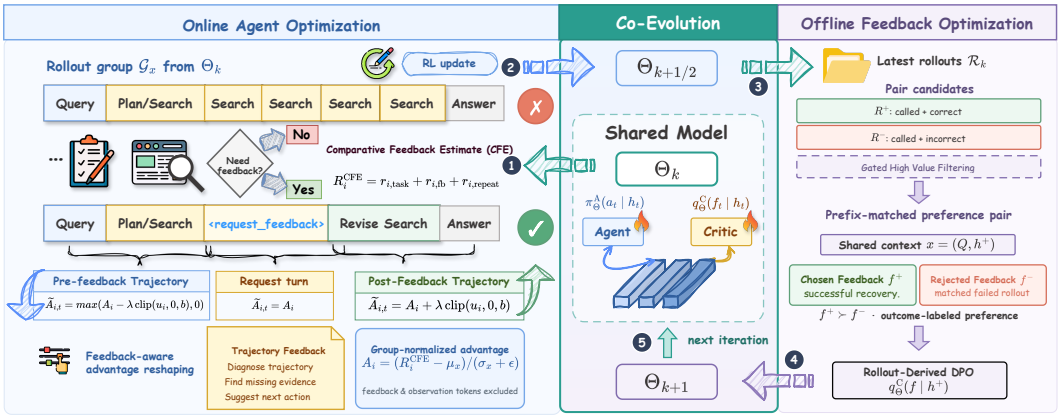


Figure 1: Overview of CAFE. One shared model serves as agent and critic. CFE shapes request returns from the call–skip gap; advantage shaping weights tokens before and after feedback. RDPO updates the critic from prefix-matched pairs mined from recent rollouts.

providing this guidance must adapt accordingly. Motivated by this coupling, we propose **CAFE**, an iterative, shared-parameter framework that integrates corrective feedback into active trajectories and alternates agent and critic optimization.

We develop CAFE around three questions: **RQ1.** (Section 2.1) How should in-trajectory feedback be structured and how can the agent learn to generate it? **RQ2.** (Section 2.2) How can an agent learn when to request feedback and how to utilize it? **RQ3.** (Section 2.3) How can a shared model co-evolve its agent and critic capabilities?

2.1 Self-Feedback Search Agent

RQ1. *How should in-trajectory feedback be structured and how can the agent learn to generate it?*

Search agents can make locally plausible choices that nevertheless steer a trajectory away from the evidence and subgoals needed for success. Such early mistakes can compound into repetitive or unproductive actions that become increasingly difficult to reverse (Zou et al., 2025; Li et al., 2025a; Wang et al., 2026b).

In-Trajectory Feedback Interaction. We therefore augment the agent's action space with an optional feedback-request action, as illustrated in Figure 1. Given a task prompt x_i, the agent interacts with the search environment to produce a rollout τ_i and receives a binary outcome reward r_i. At any intermediate history $h_{i,t}$, the agent may emit a `<request_feedback>` action. A critic then conditions on the current trajectory and generates feedback $f_{i,t}$ that identifies a corrective next step. We append the feedback to the context and return control to the agent, which continues the rollout from this augmented context.

Role-Conditioned Agent–Critic Model. In-trajectory recovery requires a critic that can diagnose the errors made by the current agent, while maintaining a separate critic incurs substantial overhead. Self-rewarding methods motivate the use of a model's own judgments as a learning signal (Yuan et al., 2024; Yang et al., 2026). Inspired by this, we use the same model to provide corrective feedback during search. We instantiate the agent and critic as role-conditioned behaviors of a single shared model:

$$a_{i,t} \sim \pi_\theta^{\mathrm{A}}(\cdot \mid h_{i,t}), \qquad f_{i,t} \sim q_\theta^{\mathrm{C}}(\cdot \mid h_{i,t}), \tag{1}$$

where A and C denote the agent and critic roles, respectively. When the agent requests feedback, the model adopts the critic role to generate $f_{i,t}$, then returns to the agent role and continues from the augmented history. The two roles remain distinct at inference time but share a common backbone.

Bootstrapping Feedback-Conditioned Search. To bootstrap both feedback use and feedback generation, we construct SFT data from the base agent's own failure trajectories. We first collect failed rollouts and ask a teacher model (Kimi-K2.5 (Team et al., 2026) in our implementation) to identify the earliest turn at which the trajectory becomes erroneous or ceases to make progress. We

preserve the agent-generated prefix through this turn, insert a <request_feedback> action, and use the teacher to generate corrective feedback together with a feedback-conditioned continuation. We retain only repaired trajectories that reach the correct final answer. Unlike pure teacher-generated demonstrations, these trajectories retain the failure patterns encountered by the base agent while providing a successful recovery from them. They therefore serve as high-quality SFT data for initializing the shared model with feedback-augmented search behavior.

2.2 Online agent optimization for Feedback Seeking and Recovery

RQ2. *How can an agent learn when to request feedback and how to utilize it?*

Long-horizon search requires credit signals that are finer grained than a terminal outcome: with feedback as an optional intervention, the outcome further reveals neither whether a request was beneficial nor how credit should be assigned around it (Wang et al., 2025; Feng et al., 2025; Zheng et al., 2025a; Zou et al., 2026). We therefore introduce feedback-aware credit assignment at two levels. Across rollouts, we use the prompt-level empirical call–skip success gap to shape the returns of feedback-requesting trajectories. Within a rollout, we redistribute credit between the behavior that preceded a request and the recovery that followed it.

Comparative Feedback Estimate (CFE) Reward. CFE compares rollouts that request feedback with those that skip it. Following GRPO (Shao et al., 2024), we sample a rollout group $\mathcal{G}_x$ for prompt x. Let $n_{\mathrm{fb},i}$ be the number of requests in rollout i and $C_i = \mathbf{1}[n_{\mathrm{fb},i} > 0]$. Rollouts with $C_i = 1$ form the call group $\mathcal{G}_{x,\mathrm{call}}$, while those with $C_i = 0$ form the skip group $\mathcal{G}_{x,\mathrm{skip}}$. If both groups are nonempty, their empirical success gap is

$$\widehat{u}(x) = \frac{1}{|\mathcal{G}_{x,\mathrm{call}}|} \sum_{j \in \mathcal{G}_{x,\mathrm{call}}} r_j - \frac{1}{|\mathcal{G}_{x,\mathrm{skip}}|} \sum_{k \in \mathcal{G}_{x,\mathrm{skip}}} r_k. \tag{2}$$

For rollout i, we assign the prompt-level statistic $u_i = \widehat{u}(x_i)$, and all rollouts for the same prompt receive the same estimate. If either route is absent, we use the batch-level fallback in Section A.3. We combine this estimate with the task outcome through three terms. The task term $r_{i,\mathrm{task}} = r_i$ retains the original answer-correctness reward. The feedback term $r_{i,\mathrm{fb}} = \beta C_i u_i$ applies the group-level gap to rollouts that request feedback. The repeat term $r_{i,\mathrm{repeat}} = -\gamma[n_{\mathrm{fb},i} - 1]_+$ penalizes only requests beyond the first, where $[z]_+ = \max(z, 0)$. The scales $\beta, \gamma \geq 0$ control the latter two terms. Their sum gives the CFE-shaped reward:

$$R_i^{\mathrm{CFE}} = r_{i,\mathrm{task}} + r_{i,\mathrm{fb}} + r_{i,\mathrm{repeat}}. \tag{3}$$

Feedback-Aware Advantage Shaping. Let μ_x and σ_x denote the mean and standard deviation of the CFE-shaped returns within $\mathcal{G}_x$. GRPO assigns every trainable agent token in rollout i the same normalized advantage:

$$A_i = \frac{R_i^{\mathrm{CFE}} - \mu_{x_i}}{\sigma_{x_i} + \epsilon}, \qquad \epsilon > 0. \tag{4}$$

A single A_i rewards the prefix that drove the search off course as much as the continuation that repaired it. This pairing is inherited from initialization: because the SFT trajectories in Section 2.1 retain the base agent's own prefix up to its earliest erroneous turn, a request tends to follow behavior that has already gone wrong. The two parts therefore play opposite roles, and reinforcing them together rewards the very behavior the agent had to abandon.

We accordingly split the trainable agent tokens of each requesting rollout ($C_i = 1$) at its first request into $\mathcal{T}_i^{\mathrm{pre}}$, $\mathcal{T}_i^{\mathrm{call}}$, and $\mathcal{T}_i^{\mathrm{post}}$, covering the tokens before the request turn, the request turn itself, and the continuation after feedback. Observation and feedback tokens are excluded from the policy loss. We bound the adjustment by clipping the prompt-level gap, $g_i = \mathrm{clip}(u_i, 0, b)$ with $b > 0$, and apply it only when $A_i > 0$ and $g_i > 0$. Eligible rollouts receive:

$$\widetilde{A}_{i,t} = \begin{cases} \max(A_i - \lambda g_i, 0), & t \in \mathcal{T}_i^{\mathrm{pre}}, \\ A_i, & t \in \mathcal{T}_i^{\mathrm{call}}, \\ A_i + \lambda g_i, & t \in \mathcal{T}_i^{\mathrm{post}}, \end{cases} \tag{5}$$

where $\lambda \geq 0$ controls the shaping strength. Together the two components answer RQ2: CFE decides across rollouts when a request is worth making, while advantage shaping decides within a rollout which behavior the resulting success should credit.

	2Wiki		HotpotQA		MuSiQue		PopQA		TriviaQA		Bamboogle		NQ		Avg.	
Method	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
Closed-source Models																
GPT-5-Mini (2025)	67.4	76.0	**59.0**	**72.0**	27.2	36.5	34.6	41.1	73.0	81.2	**62.4**	**73.2**	31.2	42.6	50.7	60.4
Gemini-2.5-Flash (2025)	63.8	73.6	54.0	67.6	24.4	35.7	40.2	49.2	73.4	81.5	57.6	65.9	39.8	**53.0**	50.5	**60.9**
Claude-4.5-haiku (2025)	61.2	68.0	51.0	63.0	17.0	22.2	36.8	43.3	66.8	73.6	52.8	62.4	32.8	42.9	45.5	53.6
Large Open-source Models																
Kimi-K2-thinking (2025)	67.8	76.2	57.4	69.9	29.4	37.8	33.4	38.9	72.0	78.5	50.4	58.7	30.2	41.1	48.7	57.3
GLM-4.7 (2025)	64.4	74.6	57.4	70.9	24.8	33.6	34.4	41.2	**73.6**	**81.6**	58.4	69.3	33.0	45.2	49.4	59.5
Qwen2.5-72B-Instruct (2025)	61.4	69.4	52.6	65.1	26.0	35.0	35.4	43.7	66.8	74.7	57.6	66.6	35.8	45.8	47.9	57.2
DeepSeek-V4-Flash-preview (2026)	59.6	66.8	51.0	62.6	17.8	23.5	22.8	26.8	61.4	67.7	47.2	52.8	28.2	38.0	41.1	48.3
RL-based Search Agent Baselines																
WebSeer† (2026)	57.8	72.3	47.8	61.3	22.6	35.1	31.6	40.8	57.8	67.3	47.2	61.2	21.6	31.9	40.9	52.8
Search-R1* (2025a)	67.0	75.4	48.4	60.9	25.8	36.2	41.0	46.9	65.0	70.8	47.2	58.4	39.8	49.1	47.7	56.8
R-Search* (2026)	69.8	77.7	52.2	64.4	31.4	**41.6**	41.8	48.1	64.2	71.7	42.4	57.6	38.0	49.1	48.5	58.6
IGPO‡ (2025)	79.6	86.1	53.4	65.1	27.8	36.9	43.2	49.1	63.2	71.4	48.8	59.1	36.8	48.1	50.4	59.4
StepSearch† (2025a)	52.6	63.2	45.2	54.5	29.2	38.8	32.2	39.1	53.2	61.7	40.0	51.2	33.6	44.1	40.9	50.4
CAFE																
Qwen2.5-7B-Instruct (2025)	44.8	54.4	41.8	53.6	20.4	28.8	34.4	42.1	56.4	65.6	37.6	49.3	31.6	41.3	38.1	47.9
+ SFT	63.2	72.4	44.2	55.0	19.0	27.5	35.2	42.1	52.6	62.5	42.4	51.8	28.8	39.1	40.8	50.1
+ SFT + GRPO	80.6	86.6	50.8	61.0	27.2	36.6	44.6	49.0	60.4	67.9	45.6	56.2	38.4	48.6	49.7	58.0
+ SFT + CAFE	**84.0**	**89.2**	53.4	64.8	30.2	39.1	**46.4**	**51.6**	62.6	69.9	50.4	61.4	**40.8**	49.2	**52.5**	60.7

Table 1: Main results on seven agentic SearchQA benchmarks. Best and second-best results are **bolded** and underlined. † Released checkpoint evaluated under our protocol. $*$ Results reported in the original paper. ‡ Our reproduction initialized from our SFT checkpoint.

2.3 Offline Feedback Optimization and Iterative Co-evolution

RQ3. *How can a shared model co-evolve its agent and feedback capabilities?*

Online policy updates continually shift the states and failures the agent encounters, so feedback learned from earlier trajectories may lose relevance. Conversely, improved feedback changes which failures the agent can recover from and thus the rollouts that drive subsequent learning. As illustrated in Figure 1, CAFE addresses this coupling by alternating online agent optimization with rollout-derived preference optimization (RDPO), which refines the critic from preference pairs mined from the latest on-policy rollouts. Because both roles share parameters, each update changes the data distribution on which the other role is refined.

Outcome-Guided Preference Filtering. To optimize the feedback side of this loop, we need to distinguish useful from ineffective guidance. Yet the environment scores only the final answer, which also depends on the preceding search and subsequent actions. We therefore construct outcome-labeled preferences from matched on-policy rollouts. At iteration k, we group the latest rollouts $\mathcal{R}_k$ by prompt and pair a successful feedback-requesting rollout with an unsuccessful one. After structural filtering, we retain pairs with similar histories at the first request and comparable feedback lengths. Under the successful history as shared context, its feedback f^+ is chosen and the feedback f^- from the matched failed rollout is rejected. The retained pairs form $\mathcal{D}_k^{\text{fb}}$. More details can be found in Section A.4.

Offline Feedback Update and Iteration. Starting from the online checkpoint $\theta_{k+\frac{1}{2}}$, we apply RDPO to $\mathcal{D}_k^{\text{fb}}$, directly preferring feedback associated with successful recovery over matched feedback from unsuccessful trajectories. Because the agent and critic roles share all parameters, RDPO updates the same full-model checkpoint and produces θ_{k+1}, rather than training a separate critic model. The next online segment then samples fresh trajectories from θ_{k+1}, from which we rebuild $\mathcal{D}_{k+1}^{\text{fb}}$ for the subsequent offline update. Through this alternating update, CAFE keeps feedback training aligned with the evolving policy while allowing improved feedback to shape the next round of on-policy experience.

3 Experiments

Dataset and Metrics. We evaluate our method on seven agentic SearchQA benchmarks. We use 2WikiMultihopQA (Ho et al., 2020) as the sole in-domain benchmark and assess out-of-domain

Online RL		2Wiki		HotpotQA		MuSiQue		PopQA		TriviaQA		Bamboogle		NQ		Avg.	
CFE	Adv. Shaping	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
✗	✗	80.6	86.6	50.8	61.0	27.2	36.6	44.6	49.0	60.4	67.9	45.6	56.2	38.4	48.6	49.7	58.0
✓	✗	83.2	88.0	52.4	63.0	28.4	36.5	45.0	48.9	60.8	68.5	47.2	56.6	38.8	47.4	50.8	58.4
✗	✓	83.2	88.2	52.4	63.2	30.6	38.5	44.6	48.9	**61.8**	**69.6**	47.2	58.9	38.6	48.2	51.2	59.4
✓	✓	**83.4**	**88.3**	**53.0**	**63.9**	**31.8**	**40.5**	**46.2**	**50.8**	61.2	69.2	**48.0**	**60.5**	**40.0**	**48.8**	**51.9**	**60.3**

Table 2: Online optimization ablation across seven benchmarks. Best and second-best results are **bolded** and underlined. **CFE** augments the task reward with the prompt-level call–skip success gap and a repeated-request cost; **Adv. Shaping** reweights pre- and post-feedback token advantages.

generalization on HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), PopQA (Mallen et al., 2023), Bamboogle (Press et al., 2023), Natural Questions (Kwiatkowski et al., 2019), and TriviaQA (Joshi et al., 2017). We report Exact Match (EM) and token-level F1 scores. The dataset characteristics, versions, and evaluation sizes are provided in Appendix A.1.

Baselines and Implementation. We compare against three groups of baselines. **(i) Closed-source LLMs.** GPT-5-Mini (Singh et al., 2025), Gemini-2.5-Flash (Comanici et al., 2025), and Claude-4.5-Haiku (Anthropic, 2025) are prompted as search agents without task-specific training. **(ii) Large open-source LLMs.** We also include Kimi-K2-Thinking (Team et al., 2025), GLM-4.7 (Zeng et al., 2025), DeepSeek-V4-Flash (Xu et al., 2026), and Qwen2.5-72B-Instruct (Qwen et al., 2025). We evaluate all models in groups (i)–(ii) under our protocol. **(iii) RL-based search agents.** We evaluate the released checkpoints of WebSeer-14B† (He et al., 2026) and StepSearch† (Zheng et al., 2025a) on our evaluation sets. Search-R1* (Jin et al., 2025a) and R-Search* (Zhao et al., 2026) use the results reported in the original papers. IGPO‡ (Wang et al., 2025) is reproduced by applying its RL procedure to our feedback-based SFT checkpoint. We use Qwen2.5-7B-Instruct (Qwen et al., 2025) as the shared backbone for both CAFE roles. Full implementation details are provided in Section A.

Main Results. Table 1 shows three main findings. (1) With a 7B backbone, CAFE achieves the highest average EM (52.5) and the second-highest average F1 (60.7) among all evaluated methods, outperforming the strongest RL-based baseline IGPO by 2.1 EM and 1.3 F1. (2) The lower block shows steady gains across training stages. Feedback-augmented SFT improves the backbone from 38.1/47.9 to 40.8/50.1 in average EM/F1, while GRPO raises the scores to 49.7/58.0. CAFE further reaches 52.5/60.7, confirming that iterative feedback optimization provides gains beyond a stronger search policy alone. (3) Compared with GRPO, CAFE improves both metrics on every benchmark, including all six out-of-domain datasets. Relative to Search-R1, its average gains are 7.4 EM and 5.9 F1 across the four multi-hop benchmarks, compared with 1.3 EM and 1.3 F1 across the three single-hop benchmarks. This gap is consistent with our motivation: in-context feedback can correct an intermediate search error before it affects the remaining retrieval steps.

At 3B scale, CAFE again improves substantially over the initial checkpoint, reaching performance comparable to several 7B search baselines (full results in Table 4). To test CAFE in a more challenging long-horizon deep-research setting, we evaluate the 7B checkpoints on BrowseComp-Plus (Chen et al., 2025). Performance improves at each training stage, with CAFE achieving the best result (Table 7), extending the same trend beyond standard SearchQA benchmarks.

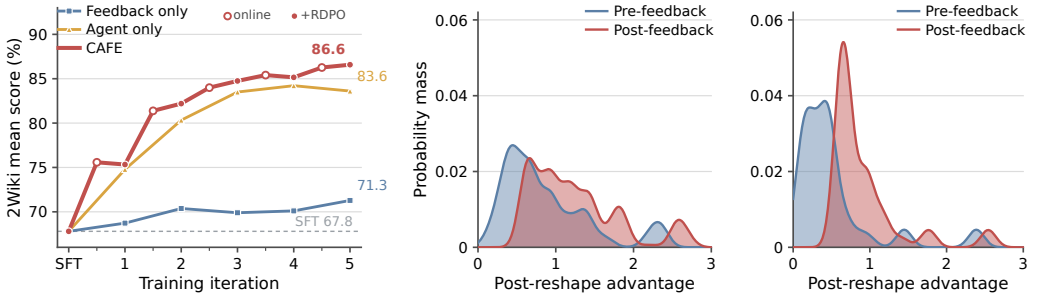
4 Ablation And Analysis

4.1 Component Ablations

Online Optimization. Table 2 compares CFE and advantage shaping under the same SFT initialization and RL budget. CFE alone raises the average EM/F1 from 49.7/58.0 to 50.8/58.4. The gain is modest but consistent with its purpose: CFE changes the return associated with requesting feedback. Advantage shaping has a larger effect, reaching 51.2/59.4 by differentially weighting agent tokens before and after feedback. This difference is also reflected in Figures 2b and 2c. The two segments begin with similar advantage distributions, but later in training the pre-feedback advantages concentrate near zero while the post-feedback advantages shift toward a positive mode. Using both components gives the best result at 51.9 EM and 60.3 F1, with improvements on every bench-

Online RL		2Wiki		HotpotQA		MuSiQue		PopQA		TriviaQA		Bamboogle		NQ		Avg.	
CFE	Adv. Shaping	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Rollout-Derived SFT (RSFT)																	
✗	✗	80.0	85.1	**54.2**	64.0	26.8	35.9	46.0	**51.8**	60.8	68.3	**50.4**	59.7	36.6	45.9	50.7	58.7
✓	✗	80.6	85.2	51.4	62.0	26.4	35.4	43.2	47.5	61.0	67.8	44.8	56.3	36.8	47.3	49.2	57.4
✗	✓	79.8	85.7	51.0	62.2	26.4	35.0	44.2	49.0	62.6	69.9	49.6	58.8	40.6	49.3	50.6	58.6
✓	✓	82.4	87.2	52.4	63.1	27.2	36.2	46.0	49.8	63.2	70.0	48.0	56.2	39.0	49.1	51.2	58.8
Rollout-Derived DPO (RDPO)																	
✗	✗	81.4	86.9	51.8	62.9	27.4	36.4	45.4	50.2	61.2	68.4	47.2	56.3	37.6	48.3	50.3	58.5
✓	✗	83.2	87.7	52.6	63.4	27.6	37.2	45.6	50.2	61.4	68.4	49.6	59.2	39.2	**49.7**	51.3	59.4
✗	✓	83.0	87.7	52.6	63.8	28.0	37.6	45.6	50.3	**64.0**	**71.5**	47.2	58.2	37.2	47.2	51.1	59.5
✓	✓	**84.0**	**89.2**	53.4	**64.8**	**30.2**	**39.1**	**46.4**	51.6	62.6	69.9	**50.4**	**61.4**	**40.8**	49.2	**52.5**	**60.7**

Table 3: Offline optimization ablation under a matched five-iteration schedule with 100 online RL steps per iteration. Best and second-best results are **bolded** and underlined. Row groups specify the offline objective (RSFT or RDPO), while the columns indicate the online RL components used to train the starting checkpoint.



(a) Agent–feedback co-evolution (b) Early training (c) Late training

Figure 2: Analysis of co-evolution and feedback-aware credit assignment. In panel (a), agent-only optimization pairs the evolving agent with the frozen SFT feedback model, whereas feedback-only optimization pairs the evolving feedback model with the frozen SFT agent. For CAFE, open markers denote intermediate online-RL checkpoints and filled markers denote the RDPO checkpoints.

mark. Full test-accuracy and empirical routing-entropy trajectories over 500 steps are reported in Section E and Fig. 5. CAFE achieves the highest late-stage test accuracy while maintaining higher feedback-policy entropy. The largest EM gain occurs on MuSiQue, while the largest F1 gain occurs on Bamboogle. Both are multi-hop benchmarks where correcting an intermediate error can affect several subsequent retrieval steps.

Offline Optimization. Learning feedback from rollout outcomes is not straightforward because the terminal label applies to the entire trajectory rather than to the feedback itself. We compare RDPO with rollout-derived SFT (RSFT), which uses the same mining pipeline as RDPO but retains only feedback from successful rollouts. This SFT objective is noisy: a trajectory may succeed because of its search prefix or subsequent actions even when the feedback is unhelpful. RDPO instead preserves the comparison with a failed rollout matched by prompt and pre-feedback state, providing a cleaner signal for feedback quality. The relative objective is also better for the shared model, since it remains anchored to the online checkpoint, whereas maximum-likelihood fitting can shift both roles without such a constraint. RDPO outperforms RSFT in three of four settings (Table 3). We perform five iterations of alternating online RL and RDPO, and each iteration contains 100 online RL steps followed by one RDPO update. More ablations are conducted in Section D.3.

4.2 Co-evolution Analysis

We compare five rounds of feedback-only, agent-only, and alternating optimization from the same feedback-augmented SFT checkpoint. In the agent-only control, online RL updates the agent while every feedback request is answered by a frozen copy of the SFT model. In the feedback-only control,

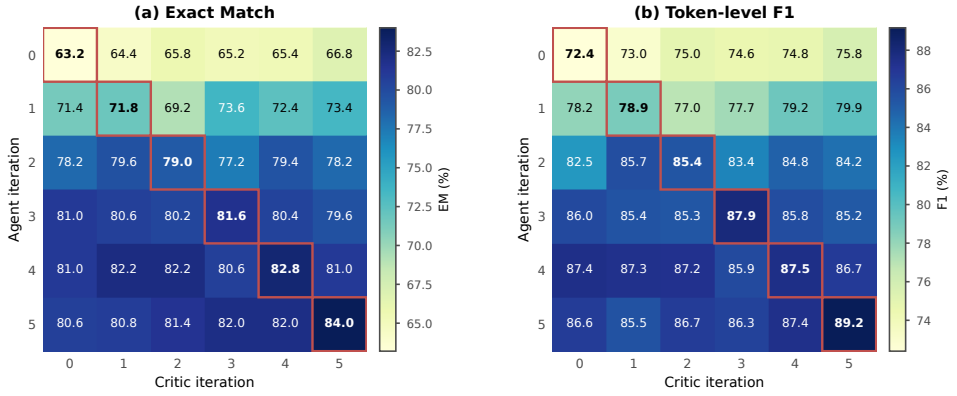


Figure 3: Agent–critic cross-play on 2Wiki. Rows and columns denote agent and critic training iterations, respectively. Iteration 0 is the shared SFT initialization. Cells report EM or token-level F1, and red boxes mark same-iteration pairs.

the SFT agent remains fixed while RDPO updates the model that generates its requested feedback. CAFE instead alternates online RL and RDPO, with the latest shared checkpoint serving both roles. Performance is measured on 2Wiki using the mean of EM and F1. As shown in Figure 2a, feedback-only optimization improves the score from 67.8 to 71.3, while agent-only optimization peaks at 84.2 before ending at 83.6. Alternating optimization reaches 86.6, outperforming the final agent-only checkpoint by 3.0 points. The one-sided controls show that improving either capability helps, while updating both allows the gains to continue across iterations.

We further test whether these gains reflect stage-specific alignment rather than a uniformly stronger critic by cross-playing agent and critic checkpoints across iterations. As shown in Figure 3, from iterations 3 through 5, each agent performs best with the critic from the same iteration. Holding the final agent fixed and replacing the SFT critic with the iteration-5 critic raises EM from 80.6 to 84.0 and F1 from 86.6 to 89.2. Conversely, the iteration-5 critic is not universally best for earlier agents, indicating that the gains arise from alignment with the policy's evolving failure distribution rather than critic strength alone.

4.3 Feedback Evolution Analysis

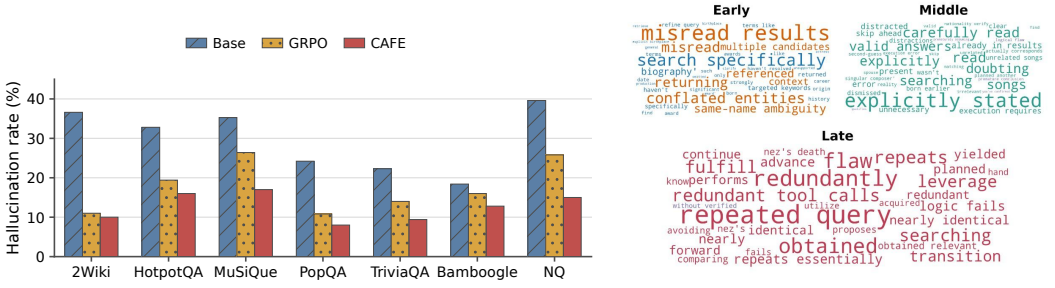
To track how feedback evolves during training, Figure 4b visualizes frequent terms from trajectories collected after iterations 1, 3, and 5, representing the early, middle, and late stages. Early feedback is dominated by retrieval and grounding errors, including ***misread results*** and ***conflated entities***. In the middle stage, the emphasis shifts toward careful evidence verification, reflected by terms such as ***valid answers***, and ***explicitly stated***. Late feedback increasingly targets residual search and reasoning inefficiencies, including ***repeated query***, ***redundant tool calls***, and ***logic fails***. This progression indicates that as basic retrieval and grounding failures recede, the critic adapts to the agent's evolving error profile by focusing increasingly on higher-level planning and execution errors.

4.4 Hallucination Analysis

Long-horizon search requires an agent to integrate evidence across many retrieval and reasoning steps, making the final answer vulnerable to unsupported claims carried forward from earlier errors. We therefore evaluate answer-level hallucination, marking an answer as hallucinated if it contains at least one factual claim unsupported by the evidence retrieved along its search trajectory. The base model produces an average hallucination rate of 29.9%, which drops to 17.6% after outcome-reward GRPO and further to 12.6% with CAFE. As shown in Figure 4a, CAFE reduces hallucinations relative to GRPO on every benchmark, with the largest reductions on NQ (10.8 percentage points) and MuSiQue (9.4 points).

5 Related Work

Search Agent and Agentic Credit Assignment. Search agents have progressed from pipelines that interleave reasoning and retrieval (Trivedi et al., 2023; Yao et al., 2023) to outcome-supervised



(a) Hallucination rate

(b) Feedback-content evolution

Figure 4: Analysis of answer grounding and feedback evolution. (a) Hallucination rates across seven benchmarks, where lower is better. (b) Dominant feedback terms during early, middle, and late training.

policies that learn when and how to search (Jin et al., 2025a; Song et al., 2025). Their long trajectories nevertheless retain a sparse-credit problem: terminal correctness does not identify which intermediate decisions were useful. Recent methods provide finer signals through information gain and confidence changes (Zheng et al., 2025a; Wang et al., 2025; Xie et al., 2026), local state comparisons and advantage shaping (Feng et al., 2025; Fan et al., 2026), diagnostic or directional signals (Zhang et al., 2026; Zou et al., 2025; 2026), and process rewards, search hints, or multi-agent refinement (Luo et al., 2025). These approaches sharpen supervision for intermediate search behavior. Once explicit feedback intervenes, however, success also couples the behavior that prompted the request with the recovery that followed it, creating a distinct feedback-conditioned credit boundary.

Self-Reflection and Corrective Feedback. Natural-language correction has been explored through prompted inference-time loops (Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2024) and through trained self-verification, self-correction, or critique models (Kumar et al., 2025; Ma et al., 2025; Xie et al., 2025). In search settings, ReSeek equips trajectories with evidence judgments and replanning (Li et al., 2025a), whereas WebSeer uses answer-submission-triggered outcome feedback to support continued search (He et al., 2026). More closely related, ECHO co-evolves separate policy and critic models through score-aware hindsight refinement, but its critic is invoked only after a trajectory is completed and therefore cannot redirect the ongoing search before errors compound (Li et al., 2026b). CAFE instead treats feedback as an optional intervention within the active trajectory.

Self-Evolving Agents. Self-evolving agents use generated experience to update task policies, supervisory signals, or system components. EvolveSearch alternates supervised fine-tuning on filtered trajectories with reinforcement learning exploration (Zhang et al., 2025a), while Self-Rewarding and EvoLM jointly improve task policies and supervisory signals (Yuan et al., 2024; Li et al., 2026a). Retroformer updates reflection and prompt revision from environmental feedback (Yao et al., 2024); AFlow, Gödel Agent, and ADAS extend optimization to agent designs, workflows, and runtime logic (Zhang et al., 2025b; Yin et al., 2025; Hu et al., 2025). These lines of work optimize different parts of the improvement loop. We study their coupling across two timescales: feedback is requested and used within a trajectory, while recent rollout outcomes update feedback generation across iterations, changing the experience available to both roles in the next round.

6 Conclusion

We introduced CAFE, a shared-model framework that jointly adapts an agent's ability to use feedback and its ability to generate it. Online RL trains the agent to request and act on feedback, while rollout-derived preference optimization updates the critic using recent trajectories. Across seven search QA benchmarks, CAFE achieves the strongest average performance among the evaluated RL-based agents, transfers to six out-of-domain datasets, and reduces answer-level hallucinations. The broader lesson is that acting and critiquing form a coupled learning system: each changes the experience from which the other improves. A self-improving agent therefore needs feedback that evolves with its policy, rather than a fixed supervisor tied to an earlier distribution of failures.

AI USE STATEMENT

We used Large Language Models (LLMs) as auxiliary tools to polish the manuscript language and improve readability. We also used generative AI to generate and curate synthetic datasets for model training, as described in Section 2.1. We conceived and executed all methodological contributions, experiments, and conclusions independently.

ETHICS STATEMENT

All authors have read and agree to adhere to the ICLR Code of Ethics. This work uses existing question-answering benchmarks and model-generated trajectories and does not involve human-subject data collection.

REPRODUCIBILITY STATEMENT

Appendix A documents the data splits, evaluation sets, backbone and hardware configuration, online and offline optimization hyperparameters, feedback-reward settings and rollout-derived preference construction used in our experiments. Appendix B provides the complete agent and critic prompts, while Appendix C gives the theoretical analysis. The remaining appendices report per-dataset ablations, training dynamics, and qualitative trajectories. We will provide a link to the corresponding code repository in the public version.

REFERENCES

Johannes Ackermann, Takashi Ishida, and Masashi Sugiyama. Off-policy corrected reward modeling for reinforcement learning from human feedback. *arXiv preprint arXiv:2507.15507*, 2025.

Kang An, Ziliang Wang, Xuhui Zheng, Faqiang Qian, Weikun Zhang, Yuhang Wang, and Wu Yichao. Erase to improve: Erasable reinforcement learning for search-augmented llms. In *International Conference on Learning Representations*, volume 2026, pp. 98392–98419, 2026.

Anthropic. Claude haiku 4.5 system card. Technical report, Anthropic, October 2025. URL https://assets.anthropic.com/m/99128ddd009bdcb/Claude-Haiku-4-5-System-Card.pdf.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent, 2025. URL https://arxiv.org/abs/2508.06600.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Wei Fan, Wenlin Yao, Zheng Li, Feng Yao, Xin Liu, Liang Qiu, Qingyu Yin, Yangqiu Song, and Bing Yin. Deepplanner: Scaling planning capability for deep research agents via advantage shaping. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 7510–7525. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.findings-acl.370/.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. *CoRR*, abs/2505.10978, 2025. doi: 10.48550/ARXIV.2505.10978. URL https://doi.org/10.48550/arXiv.2505.10978.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning*

Representations, volume 2024, pp. 57734–57811, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fef126561bbf9d4467dbb8d27334b8fe-Paper-Conference.pdf.

Guanzhong He, Zhen Yang, Jinxin Liu, Bin Xu, Lei Hou, and Juanzi Li. Webseer: Training deeper search agents through reinforcement learning with self-reflection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YCXWIfVakj.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Núria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/V1/2020.COLING-MAIN.580. URL https://doi.org/10.18653/v1/2020.coling-main.580.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=t9U3LW7JVX.

Zhengbao Jiang, Frank F 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, 2023.

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, 2025a. URL https://arxiv.org/abs/2503.09516.

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (eds.), *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pp. 737–740. ACM, 2025b. doi: 10.1145/3701716.3715313. URL https://doi.org/10.1145/3701716.3715313.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal M. P. Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=CjwERcAU7w.

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/.

Shiyu Li, Yang Tang, Yifan Wang, Peiming Li, and Xi Chen. Reseek: A self-correcting framework for search agents with instructive rewards. *CoRR*, abs/2510.00568, 2025a. doi: 10.48550/ARXIV.2510.00568. URL https://doi.org/10.48550/arXiv.2510.00568.

Shuyue Stella Li, Rui Xin, Teng Xiao, Yike Wang, Rulin Shao, Zoey Hao, Melanie Sclar, Sewoong Oh, Faeze Brahman, Pang Wei Koh, and Yulia Tsvetkov. Evolm: Self-evolving language models through co-evolved discriminative rubrics. *CoRR*, abs/2605.03871, 2026a. doi: 10.48550/ARXIV.2605.03871. URL https://doi.org/10.48550/arXiv.2605.03871.

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, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 5420–5438. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.EMNLP-MAIN.276. URL https://doi.org/10.18653/v1/2025.emnlp-main.276.

Zhicong Li, Lingjie Jiang, Yulan Hu, Xingchen Zeng, Yixia Li, Xiangwen Zhang, Guanhua Chen, Zheng Pan, Xin Li, and Yong Liu. No more stale feedback: Co-evolving critics for open-world agent learning. 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. 12643–12660, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.576. URL https://aclanthology.org/2026.acl-long.576/.

Kun Luo, Hongjin Qian, Zheng Liu, Ziyi Xia, Shitao Xiao, Siqi Bao, Jun Zhao, and Kang Liu. Infoflow: Reinforcing search agent via reward density optimization. *CoRR*, abs/2510.26575, 2025. doi: 10.48550/ARXIV.2510.26575. URL https://doi.org/10.48550/arXiv.2510.26575.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. S^2r: Teaching llms to self-verify and self-correct via reinforcement learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 22632–22654. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1104. URL https://doi.org/10.18653/v1/2025.acl-long.1104.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Vitaliy Polshkov, Marcin Pitera, Jeremy Yang, Kirill Priemko, Maksim Gaiduk, Aleksandr Nikolenko, Denis Bykov, Denis Yarats, Clare Southern, and Jerry Ma. WANDR: A benchmark for wide and deep research. Technical report, Perplexity AI, July 2026. URL https://research.perplexity.ai/articles/wandr-benchmark-evaluating-research-agents-that-must-search-wide-and-deep. Technical report. Code and tasks: https://github.com/perplexityai/wandr.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pp. 5687–5711. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.378. URL https://doi.org/10.18653/v1/2023.findings-emnlp.378.

Yunjia Qi, Zehua Yin, Xintong Shi, Hao Peng, Songyuanyi Lu, Yixian Liu, Richeng Xuan, Yuhong Liu, Zhichao Hu, Xiaozhi Wang, et al. Trajdebug: Tracing error lifecycle to identify critical failures in long-horizon agent trajectories. *arXiv preprint arXiv:2608.06346*, 2026.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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, 2025. URL https://arxiv.org/abs/2503.05592.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10: 539–554, 2022. doi: 10.1162/TACL_A_00475. URL https://doi.org/10.1162/tacl_a_00475.

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.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn LLM agents. *CoRR*, abs/2510.14967, 2025. doi: 10.48550/ARXIV.2510.14967. URL https://doi.org/10.48550/arXiv.2510.14967.

Xinyu Jessica Wang, Haoyue Bai, Yiyou Sun, Haorui Wang, Shuibai Zhang, Wenjie Hu, Mya Schroder, Bilge Mutlu, Dawn Song, and Robert D Nowak. The long-horizon task mirage? diagnosing where and why agentic systems break. *arXiv preprint arXiv:2604.11978*, 2026a.

Zehong Wang, Fang Wu, Hongru Wang, Xiangru Tang, Bolian Li, Zhenfei Yin, Yijun Ma, Yiyang Li, Weixiang Sun, Xiusi Chen, and Yanfang Ye. Why reasoning fails to plan: A planning-centric analysis of long-horizon decision making in llm agents, 2026b. URL https://arxiv.org/abs/2601.22311.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents, 2025. URL https://arxiv.org/abs/2504.12516.

Ryan Wong, Jiawei Wang, Li Chen, Yan Gao, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, Wenhao Huang, Yang Wang, et al. Widesearch: Benchmarking agentic broad info-seeking. In *International Conference on Learning Representations*, volume 2026, pp. 10012–10086, 2026.

Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu, and Weinan Zhang. A survey of llm-based deep search agents: Paradigm, optimization, evaluation, and challenges. *arXiv preprint arXiv:2508.05668*, 2025.

Yutao Xie, Nathaniel Thomas, Nicklas Hansen, Yang Fu, Li Erran Li, and Xiaolong Wang. TIPS: turn-level information-potential reward shaping for search-augmented llms. *CoRR*, abs/2603.22293, 2026. doi: 10.48550/ARXIV.2603.22293. URL https://doi.org/10.48550/arXiv.2603.22293.

Zhihui Xie, Jie Chen, Liyu Chen, Weichao Mao, Jingjing Xu, and Lingpeng Kong. Teaching language models to critique via reinforcement learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/xie25a.html.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, and Sitong Wang. SSR-zero: Simple self-rewarding reinforcement learning for machine translation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6039–6052, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.300. URL https://aclanthology.org/2026.findings-acl.300/.

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 Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

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

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh R N, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phil L Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. Retroformer: Retrospective large language agents with policy gradient optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KOZu91CzbK.

Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, and William Yang Wang. Gödel agent: A self-referential agent framework for recursively self-improvement. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pp. 27890–27913. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1354. URL https://doi.org/10.18653/v1/2025.acl-long.1354.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Dingchu Zhang, Yida Zhao, Jialong Wu, Liwen Zhang, Baixuan Li, Wenbiao Yin, Yong Jiang, Yu-Feng Li, Kewei Tu, Pengjun Xie, and Fei Huang. Evolvesearch: An iterative self-evolving search agent. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 13123–13136. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.EMNLP-MAIN.663. URL https://doi.org/10.18653/v1/2025.emnlp-main.663.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=z5uVAKwmjf.

Yaocheng Zhang, Haohuan Huang, Zijun Song, Zijie Zhao, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Criticsearch: Fine-grained credit assignment for search agents via a retrospective critic. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 12272–12290. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.findings-acl.596/.

Qingfei Zhao, Ruobing Wang, Dingling Xu, Daren Zha, Bowen Ma, Zhichun Wang, Shijie Jia, Limin Liu, and Xin Wang. R-search: Empowering LLM reasoning with search via multi-reward reinforcement learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 38030–38046. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.findings-acl.1896/.

Xuhui Zheng, Kang An, Ziliang Wang, Yuhang Wang, and Yichao Wu. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 21805–21830. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.EMNLP-MAIN.1106. URL https://doi.org/10.18653/v1/2025.emnlp-main.1106.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 414–431. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.EMNLP-MAIN.22. URL https://doi.org/10.18653/v1/2025.emnlp-main.22.

Deyu Zou, Yongqiang Chen, Jianxiang Wang, Haochen Yang, Mufei Li, James Cheng, Pan Li, and Yu Gong. T^3: Reducing belief deviation in reinforcement learning for active reasoning. *CoRR*, abs/2510.12264, 2025. doi: 10.48550/ARXIV.2510.12264. URL https://doi.org/10.48550/arXiv.2510.12264.

Deyu Zou, Yongqiang Chen, Fan Feng, Mufei Li, Pan Li, Yu Gong, and James Cheng. On information self-locking in reinforcement learning for active reasoning of LLM agents. *CoRR*, abs/2603.12109, 2026. doi: 10.48550/ARXIV.2603.12109. URL https://doi.org/10.48550/arXiv.2603.12109.

A Implementation Details

A.1 Dataset Details

Training data. For SFT, we construct a feedback-augmented bootstrap dataset of approximately 17k trajectories using the procedure described in Section 2.1. For RL, we apply a two-stage selection pipeline to a large prompt pool, yielding approximately 13k training prompts. We first sample eight rollouts per prompt through rejection sampling to form a candidate set. We then retain prompts for which at least one rollout requests feedback and reaches the correct answer, while at least one no-feedback rollout is incorrect. We regard these as high-learning-value examples near the current policy's capability boundary, as the policy succeeds along a feedback-assisted route but fails along a no-feedback route for the same prompt.

Evaluation data. We follow the evaluation protocol of R-Search (Zhao et al., 2026). For the larger multi-hop benchmarks 2WikiMultihopQA, HotpotQA, and MuSiQue, we use the test splits released by Trivedi et al. (2023) and evaluate on 500 examples per dataset. For Bamboogle, a smaller multi-hop benchmark, we use all 125 test examples provided through FlashRAG (Jin et al., 2025b). For the single-hop factoid benchmarks Natural Questions, PopQA, and TriviaQA, we use the corresponding FlashRAG test sets and randomly sample 500 examples from each dataset. All methods use the same E5 retriever over a fixed local corpus. Each search trajectory is allowed at most 30 tool calls.

A.2 Training Details

Backbone and hardware. We use Qwen2.5-7B-Instruct (Qwen et al., 2025) as the shared backbone for both the agent and critic roles. All training is conducted on 8 NVIDIA H20 GPUs, and the complete online RL stage takes approximately two days.

Online reinforcement learning. For each training prompt, we sample n_{rollout}=8 trajectories, which supply the call and skip groups for the success gap in Equation (2) and the pool from which preference pairs are mined. We use a batch size of 128, a learning rate of 1×10^{-6}, and a KL-loss coefficient of 0.001. For CFE, we set the feedback scale to $\beta = 0.5$ and the repeated-request penalty to $\gamma = 0.05$ in Equation (3). The request budget is one, so the penalty applies only from the second request onward. For token-level advantage shaping in Equation (5), we set $\lambda = 0.5$, clip the prompt-level gap at $b = 0.5$, and use a pre-feedback advantage floor of 0. Detailed prompts are provided in Section B.

Offline RDPO and iterative schedule. Each RDPO update uses a learning rate of 5×10^{-7}, a preference temperature of β=0.1, and runs for 1 epoch. Because RDPO updates the same shared parameters as the online agent, this conservative setting prevents offline preference optimization from overriding the task-solving behavior learned through RL. By default, we alternate 100 online RL steps with one RDPO update for five iterations (100×5), yielding 500 online RL steps in total. This schedule keeps feedback optimization aligned with the evolving policy while maintaining stable training. The number of retained preference pairs varies across iterations due to filtering; we cap each RDPO update at 2000 pairs.

A.3 CFE Fallback and Resolved Gap

Let $\mathcal{B}_{\text{call}} = \{j \in \mathcal{B} \mid C_j = 1\}$ and $\mathcal{B}_{\text{skip}} = \{k \in \mathcal{B} \mid C_k = 0\}$ denote the two route groups in the current rollout batch $\mathcal{B}$. When both groups are nonempty, the fallback estimate is

$$\widehat{u}_{\text{batch}} = \frac{1}{|\mathcal{B}_{\text{call}}|} \sum_{j \in \mathcal{B}_{\text{call}}} r_j - \frac{1}{|\mathcal{B}_{\text{skip}}|} \sum_{k \in \mathcal{B}_{\text{skip}}} r_k. \tag{6}$$

When the prompt group of rollout i realizes a single route, so that $\widehat{u}(x_i)$ is unavailable, the resolved estimate falls back to

$$u_i = \begin{cases} \widehat{u}_{\text{batch}}, & |\mathcal{B}_{\text{call}}| > 0 \ \wedge \ |\mathcal{B}_{\text{skip}}| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

A.4 Offline Preference Filtering and Pairing

Within each prompt bucket, we construct candidate preference pairs from called-correct and called-incorrect rollouts. For each rollout, we extract the trajectory prefix through its first closed feedback request. After removing markup and lowercasing the text, we represent each prefix h by its set of alphanumeric tokens $T(h)$ and compute token-set Jaccard similarity:

$$\mathrm{sim}(h^+, h^-) = \frac{|T(h^+) \cap T(h^-)|}{|T(h^+) \cup T(h^-)|}. \tag{8}$$

We retain pairs with $\mathrm{sim}(h^+, h^-) \geq \tau_{\mathrm{sim}}$, using $\tau_{\mathrm{sim}} = 0.7$ in all experiments. An LLM judge (GPT-5.1) then performs a second-stage quality check and removes invalid or semantically mismatched pairs. The remaining feedback pairs are used for the offline RDPO update.

B Prompt Template

Search-Agent Prompt

```
## Background Information

* You are Deep Research AI Assistant, an expert in conducting thorough, multi-step research.

The question I give you is a complex question that requires a deep research to answer.

To help you perform this task, you are equipped with one tool:
- A web search tool to help you perform search for relevant information based on the given
query.


Besides, you have a hidden environment feedback that can critique your current plan or
execution when you explicitly request it with <request_feedback></request_feedback>.
## Your Task
Do not answer the question immediately.
In the first step, you must output your plan inside <plan></plan> tags.
In later steps, you can use <tool_call></tool_call> to call tools or <answer></answer> to
provide your final answer.
When you detect that your reasoning or search process is getting stuck, becoming repetitive,
 failing to find useful evidence, or leaving you with low confidence about the next step,
you may output <request_feedback></request_feedback> to request guidance from the feedback
tool before continuing.
Even if the question appears simple, you should proactively use the feedback tool in your
reasoning process whenever it can help verify your current reasoning and reduce the risk of
an incorrect answer.
You can also re-evaluate and update your plan during the later steps.

## Output Format
You must strictly follow one and only one of the four output formats below at each step:

<think>
Your thinking process here.
</think>
<plan>
Step-by-step research plan or re-plan. Each step should be concise and action-oriented.
</plan>

or

<think>
Your thinking process here.
</think>
<tool_call>
Tool call with correct format.
```

```
</tool_call>

or

<think>
Your thinking process here.
</think>
<request_feedback>
</request_feedback>

or

<think>
Your thinking process here.
</think>
<answer>
Final answer only : a word, phrase, or number.
If it's a yes-or-no question, respond with only "yes" or "no"
No explanations or additional commentary.
</answer>

You may call one or more functions to assist with the user query.
You are provided with function signatures within <tools></tools> XML tags:
<tools>
{"type": "function", "function": {"name": "search", "description": "Search the web for
relevant information. You should use this tool if the historical search content
is not enough to answer the question. Or last search result is not relevant to the
question.", "parameters": {"type": "object", "properties": {"query": {"type": "array", "
description": "The queries to search"} }, "required": ["query"]} } }
</tools>
For each function call, return a JSON object with function name and arguments within <
tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>
```

Feedback Prompt

```
You are a trajectory critic for a Deep Research agent.

Your task is to read the user's original query and the agent's current trajectory, then
produce concise, actionable, evidence-grounded feedback that improves the agent's next step.

Requirements:
- Preserve the original task objective.
- Focus on the most important correction for the next plan or next tool call.
- Use only the information available in the provided trajectory.
- Do not solve the task.
- Do not reveal the final answer.
- Do not invent evidence that is not present in the trajectory.
- Keep the feedback brief, specific, and directly usable.
Please review the following task information:
<query> {query} </query>
<trajectory> {trajectory} </trajectory>
Evaluate how effectively the Plan addresses the Query, taking into account the real-world
feedback from the Tool Execution Trajectory.

Provide constructive, overall feedback that identifies any flaws in the logic or execution
and suggests how the plan can be improved.

Output only a single XML block in the exact format below:
<feedback>
```

```
Your constructive feedback text here
</feedback>
```

C Algorithm Analysis

C.1 CAFE Training Procedure

Algorithm 1 summarizes the complete CAFE pipeline. We first initialize the shared agent–critic model with feedback-augmented SFT data, then alternate online agent optimization with rollout-derived offline feedback optimization.

Algorithm 1 CAFE training procedure

Require: Base model θ_{base}, prompt pool $\mathcal{D}$, teacher M_T, schedule $(K, H, n, E_{\text{DPO}})$
Ensure: Co-evolved shared model θ_K

Stage I: Feedback-augmented SFT initialization

Collect base-agent failures $\mathcal{F}$ from $\mathcal{D}$; $\mathcal{D}_{\text{SFT}} \leftarrow \emptyset$
for all $\tau \in \mathcal{F}$ **do**
  $t^\star \leftarrow \textsc{LocateFirstError}(M_T, \tau)$
  Preserve $\tau_{\le t^\star}$ and insert <request_feedback>
  Let M_T generate feedback and complete the repaired trajectory τ^+
  if τ^+ reaches the correct final answer **then**
    $\mathcal{D}_{\text{SFT}} \leftarrow \mathcal{D}_{\text{SFT}} \cup \{\tau^+\}$
$\theta_0 \leftarrow \text{SFT}(\theta_{\text{base}}, \mathcal{D}_{\text{SFT}})$

Stage II: Alternating online and offline optimization

for $k = 0, \ldots, K-1$ **do**
  $\theta \leftarrow \theta_k$, $\mathcal{R}_k \leftarrow \emptyset$
  for $h = 1, \ldots, H$ **do**
    Sample rollout batch $\mathcal{B}_h$ with n trajectories per prompt and shared agent/critic role switching
    $\mathcal{R}_k \leftarrow \mathcal{R}_k \cup \mathcal{B}_h$
    Compute CFE returns R_i^{CFE} and shaped token advantages $\widetilde{A}_{i,t}$ on $\mathcal{B}_h$
    $\theta \leftarrow \text{GRPOUpdate}(\theta, \mathcal{B}_h)$
  $\theta_{k+\frac{1}{2}} \leftarrow \theta$
  $\mathcal{D}_k^{\text{fb}} \leftarrow \textsc{FilterAndPair}(\mathcal{R}_k)$ ▷ matched successful/failed feedback
  $\theta_{k+1} \leftarrow \text{RDPO}(\theta_{k+\frac{1}{2}}, \mathcal{D}_k^{\text{fb}}, E_{\text{DPO}})$
return θ_K

C.2 Preservation of the Task-Update Direction

The online CAFE update adds CFE and feedback-aware advantage shaping to the outcome-only GRPO update. We show that these terms do not reverse the original task-update direction when their induced perturbation is smaller than the baseline update norm. This ensures that learning when and how to use feedback does not optimize against task success.

For a fixed policy, let

$$\mathcal{N}(R)_i = \frac{R_i - \overline{R}}{\widehat{\sigma}(R) + \epsilon}, \quad q_i = \beta C_i u_i - \gamma[n_{\text{fb},i} - 1]_+, \quad A_i^0 = \mathcal{N}(r)_i, \quad A_i = \mathcal{N}(r+q)_i,$$

where $\widehat{\sigma}$ is the population standard deviation within the rollout group, and let $\widetilde{A}_{i,t}$ be the final shaped token advantage. For $s_{i,t} = \nabla_\theta \log \pi_\theta(a_{i,t} \mid h_{i,t})$ and common token weights $w_{i,t} \ge 0$, define the outcome-only task update and the online CAFE update as

$$G_{\text{task}} = \mathbb{E}\left[\sum_{i,t} w_{i,t} s_{i,t} A_i^0\right], \qquad G_{\text{CAFE}} = \mathbb{E}\left[\sum_{i,t} w_{i,t} s_{i,t} \widetilde{A}_{i,t}\right].$$

Assume $r_i \in [0,1]$, $|u_i| \le U$, $[n_{\text{fb},i} - 1]_+ \le K$, $0 \le g_i \le b$, and that both normalization denominators are at least $\nu > 0$. Reward statistics and shaping coefficients are treated as stop-gradient quantities. We further assume $B_\pi := \sup_{\mathcal{G}} \sum_{i,t} w_{i,t} \|s_{i,t}\| < \infty$, where $\mathcal{G}$ denotes a rollout group.

Theorem 1 (Task-update direction preservation). *Let* $\eta_R = \beta U + \gamma K$ *and* $c_{\mathrm{norm}} = 2/\nu + 1/\nu^2$, *and define*

$$\Delta_{\mathrm{CAFE}} = B_\pi[c_{\mathrm{norm}}\eta_R + \lambda b]\,.$$

If $\Delta_{\mathrm{CAFE}} < \|G_{\mathrm{task}}\|$, *then*

$$\langle G_{\mathrm{CAFE}}, G_{\mathrm{task}}\rangle \geq \|G_{\mathrm{task}}\|\,(\|G_{\mathrm{task}}\| - \Delta_{\mathrm{CAFE}}) > 0.$$

Thus, the online CAFE update remains positively aligned with the original outcome-only task update.

Proof. Since $|q_i| \leq \eta_R$, we have $|q_i - \overline{q}| \leq 2\eta_R$ and $|\widehat{\sigma}(r+q) - \widehat{\sigma}(r)| \leq \eta_R$. The denominator bound and $|r_i - \overline{r}| \leq 1$ therefore imply

$$|A_i - A_i^0| \leq c_{\mathrm{norm}}\eta_R.$$

Advantage shaping changes any eligible pre- or post-feedback token by at most λb, while leaving other tokens unchanged. Hence $|\widetilde{A}_{i,t} - A_i^0| \leq c_{\mathrm{norm}}\eta_R + \lambda b$. Multiplying by the policy scores and applying the definition of B_π gives

$$\|G_{\mathrm{CAFE}} - G_{\mathrm{task}}\| \leq \Delta_{\mathrm{CAFE}}.$$

Therefore, by Cauchy–Schwarz,

$$\begin{aligned}\langle G_{\mathrm{CAFE}}, G_{\mathrm{task}}\rangle &= \|G_{\mathrm{task}}\|^2 + \langle G_{\mathrm{CAFE}} - G_{\mathrm{task}}, G_{\mathrm{task}}\rangle \\ &\geq \|G_{\mathrm{task}}\|^2 - \Delta_{\mathrm{CAFE}}\|G_{\mathrm{task}}\|,\end{aligned}$$

which is positive under the stated condition.

D Additional Ablation Results

D.1 Model Size Ablation

To test whether CAFE's gains depend on the capacity of the 7B backbone, we repeat the training pipeline with Qwen2.5-3B-Instruct (Qwen et al., 2025) under the same settings and compare it with existing 3B search agents. Table 4 reports the 3B results on all seven benchmarks; the corresponding 7B results are given in Table 1. At 3B scale, CAFE reaches an average EM/F1 of 48.8/57.4, outperforming both existing 3B baselines by a clear margin. Its performance is also comparable to several 7B search agents, indicating that the benefit of coupled feedback learning is not confined to higher-capacity backbones.

Method	**2Wiki**		**HotpotQA**		**MuSiQue**		**PopQA**		**TriviaQA**		**Bamboogle**		**NQ**		**Avg.**	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
Existing 3B Search Agents																
Search-R1-3B* (2025a)	58.8	68.1	46.2	57.8	24.4	32.9	37.0	43.5	56.6	63.2	41.6	53.9	34.4	44.1	42.7	51.9
R-Search-3B* (2026)	65.0	72.6	43.4	54.4	<u>25.8</u>	34.8	37.0	44.9	56.0	64.0	37.6	49.8	35.2	<u>46.0</u>	42.9	52.4
CAFE (Qwen2.5-3B-Instruct)																
Qwen2.5-3B-Instruct	11.4	27.6	14.6	25.8	3.8	9.2	11.6	18.2	18.8	32.1	12.8	22.9	3.8	12.8	11.0	21.2
+ SFT	56.2	66.8	36.0	45.2	16.6	25.4	30.6	38.2	45.2	54.5	32.0	44.3	27.0	38.1	34.8	44.6
+ SFT + GRPO	<u>78.4</u>	<u>83.8</u>	<u>47.6</u>	<u>58.2</u>	25.2	<u>35.3</u>	<u>40.0</u>	<u>46.6</u>	<u>59.6</u>	<u>67.1</u>	<u>43.2</u>	<u>54.4</u>	<u>35.8</u>	45.0	<u>47.1</u>	<u>55.8</u>
+ SFT + CAFE	**80.2**	**86.1**	**49.4**	**59.9**	**26.2**	**35.8**	**42.4**	**47.6**	**61.6**	**68.9**	**44.0**	**56.3**	**37.6**	**47.2**	**48.8**	**57.4**

Table 4: Results for 3B-scale models. Best and second-best results are bolded and underlined, respectively.

D.2 Detailed Hallucination Results

We report the per-dataset hallucination rates underlying Figure 4a. We use the answer-level criterion defined in Section 4.4 and evaluate on the same seven test sets as the main results. As shown in Table 5, outcome-reward GRPO reduces the average hallucination rate from 29.88% to 17.63%,

while CAFE further lowers it to 12.60%. CAFE improves over GRPO on every benchmark, with the largest reductions on NQ (10.8 percentage points) and MuSiQue (9.4 points).

Method	2Wiki	HotpotQA	MuSiQue	PopQA	TriviaQA	Bamboogle	NQ	Avg.
Base	36.60	32.80	35.27	24.20	22.29	18.40	39.60	29.88
GRPO	<u>11.00</u>	<u>19.40</u>	<u>26.40</u>	<u>10.82</u>	<u>14.00</u>	<u>16.00</u>	<u>25.80</u>	<u>17.63</u>
CAFE	**10.00**	**16.00**	**17.00**	**8.00**	**9.40**	**12.80**	**15.00**	**12.60**

Table 5: Hallucination rates (%, lower is better) across seven benchmarks. Best and second-best results are bolded and underlined, respectively.

D.3 Iteration Schedule Ablation

Table 6 compares three online–offline schedules under the same budget of 500 online RL steps. We denote a schedule by $H \times K$, where H online RL steps are followed by one RDPO update and the cycle is repeated for K iterations. Among the tested schedules, 100×5 achieves the highest average EM and F1 and performs best on most datasets.

	2Wiki		HotpotQA		MuSiQue		PopQA		TriviaQA		Bamboogle		NQ		Avg.	
Schedule	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
50×10	81.4	85.5	<u>51.6</u>	62.8	27.6	<u>37.4</u>	<u>46.0</u>	<u>50.9</u>	58.6	<u>67.1</u>	<u>48.8</u>	<u>58.2</u>	34.0	43.8	49.7	<u>58.0</u>
$\mathbf{100 \times 5}$	**84.0**	**89.2**	**53.4**	**64.8**	**30.2**	**39.1**	**46.4**	**51.6**	**62.6**	**69.9**	**50.4**	**61.4**	**40.8**	**49.2**	**52.5**	**60.7**
250×2	<u>81.6</u>	<u>86.5</u>	**53.4**	<u>63.1</u>	<u>28.2</u>	<u>37.4</u>	44.0	47.6	<u>59.4</u>	66.3	45.6	57.3	<u>36.8</u>	<u>46.4</u>	<u>49.9</u>	57.8

Table 6: Iteration-schedule ablation under 500 online RL steps. A schedule $H \times K$ performs H online RL steps followed by one RDPO update and repeats this cycle K times. Best and second-best results are bolded and underlined, respectively.

D.4 Results on BrowseComp-Plus

BrowseComp and BrowseComp-Plus. BrowseComp (Wei et al., 2025) evaluates deep-research agents on short-answer questions whose solutions require persistent browsing for hard-to-find and interconnected evidence. Although its live, black-box search API reflects realistic browsing conditions, the dynamic backend limits controlled and reproducible evaluation. BrowseComp-Plus (Chen et al., 2025) addresses this issue by replacing live search with a fixed curated corpus and a shared local retriever built from human-verified supporting documents and mined hard negatives, enabling consistent comparison under a controlled retrieval environment.

Table 7: EM and F1 scores (%) on BrowseComp-Plus.

Setting	EM	F1
Qwen2.5-7B-Instruct	4.4	6.8
+ SFT	5.3	7.9
+ SFT + GRPO	6.8	9.9
+ SFT +**CAFE**	7.7	10.6

To examine whether the learned feedback mechanism extends to realistic long-horizon deep-research tasks, we also evaluate the 7B checkpoints from our main experiments on BrowseComp-Plus (Chen et al., 2025). As shown in Table 7, performance improves steadily across training stages. The improvement suggests that trajectory-level feedback remains useful when solving substantially longer and more demanding search tasks.

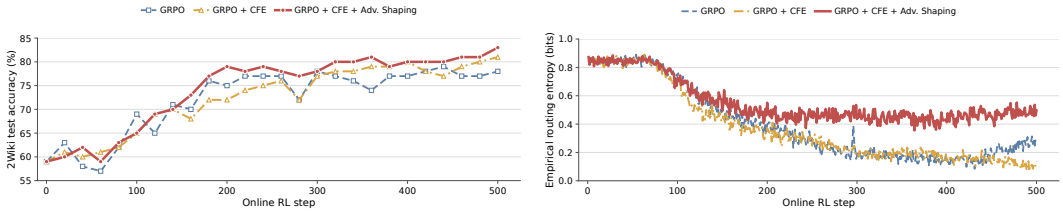
E Training Dynamics

Figure 5 compares test-accuracy and feedback-policy-entropy dynamics under the same 500 online RL steps. All test-accuracy curves include the shared SFT checkpoint at step 0 and report evaluations every 20 steps. Outcome-only GRPO improves rapidly but fluctuates after roughly 200 steps. CFE raises late-stage test accuracy, while adding feedback-aware advantage shaping yields the highest final value and a more sustained improvement.

We measure prompt-level empirical feedback-routing entropy. For prompt x, let p_x be the fraction of rollouts that request feedback and define

$$H_x = -p_x \log_2 p_x - (1 - p_x) \log_2(1 - p_x),$$

with $0 \log_2 0 = 0$. We report the mean of H_x over prompts. At the final checkpoint, CAFE retains 0.496 bits of routing entropy, compared with 0.221 for GRPO and 0.099 for GRPO+CFE. The substantially higher entropy of CAFE indicates that advantage shaping prevents premature routing collapse and preserves the ability to request feedback selectively as trajectories enter different states. Since CAFE also achieves higher test accuracy, this diversity reflects useful exploration and adaptive feedback routing rather than collapsing into a fixed pattern.



(a) Test accuracy

(b) Empirical feedback-routing entropy

Figure 5: Online training dynamics through 500 RL steps. (a) Test accuracy on 2Wiki. (b) Prompt-level empirical entropy of feedback policy.

F LIMITATIONS

Given the cost and budget required for controlled live-web evaluation, our experiments focus on fixed local retrieval corpora and do not yet cover dynamic live-web environments. Feedback requests also introduce additional inference cost, although CAFE learns to invoke them selectively. Future work will study live-web deployment and more adaptive feedback budgets.

G CASE STUDY

G.1 CAFE: FEEDBACK-CONDITIONED RECOVERY

Dataset. BrowseComp-Plus **CAFE RECOVERY**

Query. I am looking for the name of a trail about 0.50–1 mile in length and 1–3 feet in width with an elevation gain of about 150–400 feet. The trail includes a structure dating back to the 1800s. As of December 2022, it is located about 218–220 miles in aerial distance from an airport in Colorado and 1,104–1,106 miles in aerial distance from an airport in Chicago.

Ground truth. Old Kiln Trail



1. Initial airport-distance search

```
<think>
To find the name of this specific trail, I need to first identify the
location based on the given information. The trail is located between
218-220 miles in aerial distance from an airport in Colorado and
1,104-1,106 miles in aerial distance from an airport in Chicago. This
suggests the trail is in the Rocky Mountains region, likely in
Colorado or nearby states.
</think>

<plan>
1. Search for the location of the trail based on the given aerial
   distances.
2. Once the location is identified, search for the name of the trail
   that matches the described characteristics.
</plan>

<think>
I will start by searching for the location of the trail based on the
given aerial distances from the airports in Colorado and Chicago.
</think>

<tool_call>
{"name": "search", "arguments": {"query":
["Location of the trail based on aerial distances from airports in
Colorado and Chicago"]}}
</tool_call>

<tool_response>
[Result 1]
Title: The distance from Chicago O'Hare International Airport to
Grand Junction Regional Airport is:
Date: 2025-01-01

1,100 miles / 1 771 km flying

The total driving distance from ORD to GJT is 1,239 miles or 1 994
kilometers.

The total straight line flight distance from ORD to GJT is 1,100
miles. This is equivalent to 1 771 kilometers or 956 nautical miles.

Your trip begins at Chicago O'Hare International Airport in Chicago,
Illinois. It ends at Grand Junction Regional Airport in Grand
Junction, Colorado.

Your flight direction from ORD to GJT is West (-93 degrees from
North).

The distance calculator helps you figure out how far it is to get from
```

```
ORD to GJT. It does this by computing the straight line flying
distance ("as the crow flies") and the driving distance if the route
is drivable. It uses all this data to compute the total travel
mileage.

ICAO: KORD
FAA: ORD
City: Chicago
State: Illinois
Zip code: 60666
Country: United States
Category: airports

ICAO: KGJT
FAA: GJT
City: Grand Junction
State: Colorado
Zip code: 81506
Country: United States
Category: airports

Travelmath helps you find distances based on actual road trip
directions, or the straight line flight distance. You can get the
distance between cities, airports, states, countries, or zip codes to
figure out the best route to travel to your destination. Compare the
results to the straight line distance to determine whether it is
better to drive or fly. The database uses the latitude and longitude
of each location to calculate distance using the great circle distance
formula. The calculation is done using the Vincenty algorithm and the
WGS84 ellipsoid model of the Earth, which is the same one used by most
GPS receivers. This gives you the flying distance "as the crow flies."

[Result 2]
Title: O'Hare International Airport - Wikipedia
Date: 2003-06-18

Name: Chicago O'Hare International Airport
IATA: ORD
ICAO: KORD
FAA: ORD
Type: Public
Owner-operator: Chicago Department of Aviation
City served: Chicago metropolitan area
Location: O'Hare, Chicago, Illinois, U.S.
Opened: 1944
Timezone: CST
UTC: UTC-06:00
Summer: CDT
UTCS: UTC-05:00
Elevation: 668 feet / 204 meters
Coordinates: 41 58 43, 87 54 17

[Result 3]
Title: Simple Flying
Author: Jake Hardiman
Date: 2022-09-03

Denver International Airport (DEN) is an intercontinental hub situated
in the US state of Colorado. It stands out as being North America's
largest airport by surface area, and second in the world only to King
Fahd International Airport (DMM) in Dammam, Saudi Arabia. One reason
for its large surface area is its extremely long runways, but why does
it need these?

Denver International Airport opened in February 1995, replacing the
city's former Stapleton International Airport. Its first five runways
were equal in length. Runways 7/25, 8/26, 16L/34R, 17L/35R, and
17R/35L all measured exactly 12,000 feet long. This distance equates
to 3,658 meters, or 2.27 miles. However, they were eventually dwarfed
by another landing strip.

September 2003 saw the opening of Denver's sixth runway, known as
```

```
16R/34L. This concrete strip differed from the existing five runways
by being 4,000 feet longer than them. This gave it a total length of
16,000 feet, equivalent to 4,877 meters or 3.03 miles. This
astronomically long runway represented a 33% extension compared to the
five that had come before it.

The reason that Denver's runways are so long is down to the altitude
at which the airport is situated. Colorado's capital is known as the
"Mile High City," and, correspondingly, its airport finds itself
situated 1,656 m / 5,434 ft above sea level. This results in a
challenging set of atmospheric conditions for pilots known as "hot and
high." Low air density is a characteristic of hot and high conditions.
[Search excerpt truncated.]
</tool_response>
```

Stage 1 analysis: the results point to Grand Junction

The first result reports an O'Hare–Grand Junction distance of roughly 1,100 miles and identifies Grand Junction as a location in Colorado. This is a plausible geographic lead, but the agent does not use it to formulate a trail query that includes the remaining attributes.

2. Missed geographic anchor and feedback intervention

```
<think>
Wait, I'm getting confused by the search results. The query mentions
Colorado airports, but I'm seeing information about Chicago O'Hare and
Denver International Airport. I need to re-read these results more
carefully - they actually do mention Colorado airports like Grand
Junction Regional Airport (KGJT). I should focus on extracting the
geographic information from what I already have rather than searching
again with similar queries.
</think>

<request_feedback>
</request_feedback>

<feedback>
Core objective: Identify the name of a specific trail based on its
location characteristics. Issue: You misread the search results - the
tool response actually contains geographic information about Colorado
airports (Grand Junction Regional Airport, etc.) that is relevant to
your query. You incorrectly concluded the search failed when the
information was present. Fix: Re-examine the existing search results
carefully, noting the specific Colorado locations mentioned (Grand
Junction, Colorado; Denver, Colorado). Use this geographic context to
search specifically for trails in the Rocky Mountains region of
Colorado that match the described characteristics (length, width,
elevation gain, and the 1800s-era structure).
</feedback>

<think>
You're right, I misread the search results. The results do mention
Colorado airports like Grand Junction Regional Airport (KGJT) and
Denver International Airport (DEN). I need to search for trails in the
Rocky Mountains region of Colorado that match the described
characteristics. Let me search specifically for trails in Colorado
with the given attributes.
</think>
```

Stage 2 analysis: feedback turns the diagnosis into a search query

Before requesting feedback, the agent recognizes that it has misread the results but does not specify how to revise the search. The critic points the agent back to the Colorado locations in the existing results and recommends combining that context with the trail's length, width, elevation gain, and nineteenth-century structure. The feedback provides a concrete next step rather than a new diagnosis.

3. Search combining geography and trail attributes

```
<tool_call>
{"name": "search", "arguments": {"query":
["Colorado Rocky Mountains trail 0.50-1 mile 1-3 feet wide 150-400
feet elevation gain 1800s structure"]}}
</tool_call>

<tool_response>
[Result 1]
Title: Old Kiln Trail
Date: 2009-12-01

Round Trip Distance: 1 mile
Difficulty: Easy
Elevation: 6234 - 6302 feet
Cellphone: 3-5 bars
Time: 30 mins.
Trailhead: Old Kiln
Fee: none
Attractions: Old lime kiln

The Old Kiln Trail is located in the BLM Bangs Canyon Management Area,
just off of Little Park Road, south of Grand Junction. The trail
follows a former dirt road that passes near an old primitive lime
kiln. Lime kilns were used to heat limestone to make quicklime. The
resulting product had a variety of uses including making mortar and
whitewash.
If you are driving up Little Park Road it is hard to miss the
trailhead which is just past the turnoff for the Bangs Canyon Staging
Area where there are restrooms and the trailheads for the Mica Mine
and Rough Canyon trails.
After passing through the fence follow the old road to the left and
follow it as it travels northwesterly toward the monument.
There is a fork to the right in the road just before you go a quarter
mile. You will be able to see the kiln behind the trees about a
hundred yards away.
The kiln was constructed with sandstone rocks that were probably
quarried in the area and sealed with bentonite which is also
plentiful. The limestone would have been reduced to quicklime by
layering it inside the kiln with wood or coal. The layered stack would
have then been set on fire from the bottom and allowed to burn. The
lime would begin to break down when the temperature reached 900 C
(1652 F).

This was probably a vent hole to allow the fire to breathe. This small
inefficient kiln probably became obsolete when the railroad or other
means of transportation were able to supply the valley with what it
needed.

There are some nice views of the surrounding area from the hill above
the kiln. A walk out to the end of the road provides a nice overlook
of the Ribbon Trail area.
[Search excerpt truncated.]

[Result 2]
Title: Old Kiln
Date: 2025-01-01

Difficulty: Medium
Length: 0.80 Miles
Trail Width: 1-3 feet
Elevation Gain: 153
Activity: Hikers, Dogs, Horses
Trail Description:
The Old Kiln Trail is a short trail that travels through open fields,
cliffsides, ridges and rock outcroppings. Enjoy this trail's unique
geological features of landslides, faults, fossils, fish scales and
mudstone layers. Keep an eye out for the historic limestone kiln
dating from the 1800s. Dogs must be on leash.

[Result 3]
```

```
Title: Colorado Hundred Highest / Centennials
Date: 2020-08-

Colorado Hundred Highest Mountains / Centennials
Eric Gilbertson
June 16 - July 19, 2020
New Fastest Known Time: 33 days 23 hours (self-supported, incorrectly
listed as "supported" on FKT website)
Peaks: 100 (plus 20 bonus peaks)
Miles hiked: 685
Elevation gain: 295,000 ft
Rest days: 1
Longest day: 49 miles, 18k ft gain
Most Centennials in a day: 10
The Colorado Centennials are the hundred tallest mountains in
Colorado with at least 300 ft of prominence. They range in elevation
from 13,809 ft (Dallas Peak) to 14,433 ft (Mt Elbert). Many are
standard hikes and scrambles, but three peaks are technical (Dallas,
Teakettle, and Jagged).
[Search excerpt truncated.]
</tool_response>
<think>
Perfect! I found the answer. The Old Kiln Trail matches all the
characteristics described in the question:
- Length: 1 mile (also mentions 0.80 Miles)
- Width: 1-3 feet
- Elevation Gain: 153 feet
- Location: Grand Junction, Colorado (specifically in the BLM Bangs
  Canyon Management Area, just off of Little Park Road)
- 1800s-era structure: The historic limestone kiln

The search results clearly state "Keep an eye out for the historic
limestone kiln dating from the 1800s" and describe the kiln's
construction details. This is the specific trail that matches all the
given criteria.
</think>

<answer>
Old Kiln Trail
</answer>
```

Stage 3 analysis: the revised query identifies the trail

The revised query restricts the search to Colorado's Rocky Mountains and includes all four trail attributes. It retrieves *Old Kiln Trail*, whose reported attributes match the query: a length of 0.80 miles, a width of 1–3 feet, 153 feet of elevation gain, and a limestone kiln dating to the 1800s. The agent then returns the correct answer.

Case analysis: what changes after feedback

Role of feedback. The initial search result reports an ORD–GJT distance of roughly $1{,}100$ miles and identifies GJT as Grand Junction, Colorado. This is a strong geographic lead, but the agent does not use it to formulate the next trail query. In the BASE rollout, the agent instead commits to Denver and eventually returns *Old Elitch's Park Trail*.

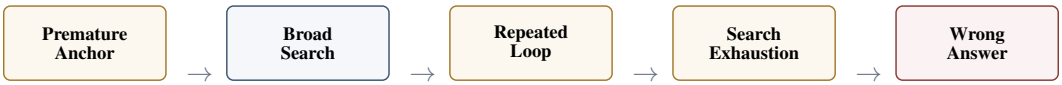
Before requesting feedback, the CAFE agent already recognizes that it has misread the result, so the critic does not supply a new diagnosis. Instead, the critic directs the agent back to the retrieved geographic evidence and specifies which trail attributes to search together. The resulting Colorado-focused query retrieves *Old Kiln Trail* near Grand Junction, along with all four matching attributes, and leads to the correct answer (EM $= 1$). Here, feedback is useful because it turns recognition of the error into a specific next query. The trajectory does not, however, verify both airport distances independently; the recovery relies on the Colorado location constraint and the four trail attributes.

G.2 BASE: UNRECOVERED SEARCH DRIFT

Dataset. BrowseComp-Plus **BASE FAILURE**

Shared query. The same trail-identification query shown above.

Ground truth. Old Kiln Trail **BASE output.** Old Elitch's Park Trail



1. Initial search under an unverified Denver assumption

```
<think>
To find the trail that matches the given criteria, I first need to identify the location. The trail is
near an airport that is 218-220 miles from a Colorado airport and 1,104-1,106 miles from a Chicago
airport. This narrows down the possible location to the vicinity of Denver, Colorado, as it's the
closest major airport to both the specified distances.
</think>
<plan>
1. Search for airports in Colorado that are approximately 218-220 miles from Denver.
2. Search for airports in Illinois that are approximately 1,104-1,106 miles from Denver.
3. Identify the trail that matches the length, width, elevation gain, and includes a 19th-century
structure.
</plan>

<think>
Based on the distances provided, the most likely airport in Colorado is Denver International Airport (
DEN). The airport in Illinois that is approximately 1,104-1,106 miles from Denver could be O'Hare
International Airport (ORD).
</think>
<tool_call>
{"name": "search", "arguments": {"query": ["trails near Denver, Colorado with 19th-century structure", "
trails near Chicago, Illinois with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails (Colorado Springs), the Old Spanish
Trail,
Trail Distances, and Lincoln Home National Historic Site. No result matched all trail constraints
jointly.]
</tool_response>
<think>Based on the search results, the trail near Denver, Colorado that includes a structure from the
19th century is the Old Spanish Trail. However, the Old Spanish Trail is a historical route that spans
multiple states and is not a specific local trail. For the trail near Chicago, Illinois, the Lincoln
Home National Historic Site includes walking paths that are on a paved surface and surrounded by various
 plant and tree species, but it does not mention a specific 19th-century structure. I need to search for
 more specific trails near Denver that might fit the criteria better.</think>
```

Stage 1 analysis: Denver is assumed rather than verified

The agent infers Denver before establishing which airports define the two distances. Its broad query over Denver and Chicago returns no trail that satisfies the full set of constraints. Nevertheless, the agent retains Denver and treats the Old Spanish Trail as a candidate, even though that route does not match the requested trail. The failure begins with an unverified location assumption, not retrieval alone.

2a. Early Denver-specific requery

```
<tool_call>
{"name": "search", "arguments": {"query": ["trails near Denver, Colorado with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails (Colorado Springs), the Old Spanish
Trail,
and Trail Distances. None matched the required location, dimensions, elevation gain, and nineteenth-
century structure jointly.]
```

```
</tool_response>
<think>After further searching, there is no specific trail in Denver, Colorado that matches the exact
criteria of being about 0.50 - 1 mile in length, 1 - 3 feet in width, and having an elevation gain of
150 - 400 feet with a 19th-century structure. The Old Spanish Trail is a historic route, but it is much
longer and not a local trail. The trails near Denver, such as the Urban Trails, do not seem to fit the
specific length and width criteria, and none of them mention a 19th-century structure. I will now search
 for trails near the vicinity of the airport in Colorado, which is likely Denver International Airport,
to find a trail that might match the criteria better.</think>
<tool_call>
<tool_response>
[Search results summarized for readability: Returned the Old Spanish Trail, a Denver airport runway
article,
and Colorado high-peak material. These reinforced the Denver/Colorado frame but did not identify a trail
 satisfying the joint constraints.]
</tool_response>
<think>From the search results, I still haven't found a trail that matches the specific criteria of
being about 0.50 - 1 mile in length, 1 - 3 feet in width, and having an elevation gain of 150 - 400 feet
 with a 19th-century structure. The Old Spanish Trail is a historic route, but it is much longer and not
 a local trail. The information provided about Denver International Airport and Colorado's highest
mountains does not seem to be directly relevant to the trail search. I need to narrow down the search to
 trails near Denver, Colorado, that might have a 19th-century structure and fit the length and elevation
 gain criteria. I will search for trails near Denver that might have historical structures or are part
of the broader network of trails that could potentially fit the description.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["trails near Denver, Colorado with 19th-century structure and
 elevation gain 150-400 feet"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Leadville mining-town and high-elevation-city pages
 plus
Trail Distances. They supplied Colorado history and trail lengths, but no single trail matched the full
constraint set.]
</tool_response>
<think>Based on the search results, there is no specific trail near Denver, Colorado that matches the
exact criteria of being about 0.50 - 1 mile in length, 1 - 3 feet in width, and having an elevation gain
 of 150 - 400 feet with a 19th-century structure. The information provided about Leadville, Colorado,
highlights its rich history and the presence of historic buildings, but it does not mention a trail that
 fits the given criteria. I will now focus on trails near Denver that are known to have historical
structures and fit the length and elevation gain criteria.</think>
<tool_call>
Tool call with correct format.
<tool_call>
{"name": "search", "arguments": {"query": ["trails near Denver, Colorado with 19th-century structure and
 elevation gain 150-400 feet"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Repeated the same Leadville pages and Trail Distances
material.
No result identified a Denver-area trail matching the location, dimensions, elevation gain, and
nineteenth-century structure together.]
</tool_response>
<think>After further searching, I still haven't found a trail near Denver, Colorado that matches the
specific criteria of being about 0.50 - 1 mile in length, 1 - 3 feet in width, and having an elevation
gain of 150 - 400 feet with a 19th-century structure. The information provided about Leadville, Colorado
, highlights its rich history and the presence of historic buildings, but it does not mention a trail
that fits the given criteria. The trails near Denver, such as the Urban Trails, do not seem to have 19th
-century structures and do not fit the length and elevation gain criteria. It is possible that the trail
 I am looking for is not well-documented or might be a local path that is not listed in the general
trail databases. I will try to search for more specific local sources or historical records that might
mention such a trail.</think>
```

Stage 2a analysis: repeated failures leave the Denver assumption unchanged

The agent adds historical structures and elevation gain to the query but continues to restrict every search to Denver. When these queries return no match, it does not reconsider the location. Instead, each failure prompts another Denver-specific search.

2b. Historical and local sources, still centered on Denver

<tool_call>

<tool_call>
{"name": "search", "arguments": {"query": ["local trails in Denver, Colorado with 19th-century structure and elevation gain 150-400 feet"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Trail Distances, Leadville history, and Colorado high-peak material. None identified a local Denver trail satisfying all requested attributes.]
</tool_response>
<think>After searching local trails in Denver, Colorado, for a trail with a 19th-century structure, a length of about 0.50 - 1 mile, a width of 1 - 3 feet, and an elevation gain of 150 - 400 feet, I still haven't found a suitable trail. The search results do not provide any specific trail that matches all these criteria. It is possible that the trail I am looking for is not well-documented or might be a local path that is not listed in the general trail databases. Given the lack of relevant information, I will re-evaluate the plan and consider searching for more specific historical records or local community knowledge to find the trail.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["historical trails in Denver, Colorado with 19th-century structure and elevation gain 150-400 feet"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Trail Distances, a Leadville feature, and Mesa Verde
and Lewis and Clark historical-park pages. No result matched the joint trail constraints.]
</tool_response>
<think>Despite searching for historical trails in Denver, Colorado with a 19th-century structure, an elevation gain of 150-400 feet, and a length of about 0.50 - 1 mile, the search results have not provided a specific trail that matches all these criteria. The information provided about Leadville, Colorado, highlights its rich historical and mining boomtown heritage, but does not mention a trail that fits the given criteria. The historical sites mentioned in the search results, such as Mesa Verde National Park and Lewis and Clark National Historical Park, are not located near Denver and do not seem to have trails that match the specified criteria. Given the lack of relevant information, I will consider searching for local historical societies or city records that might have more detailed information about the trails in the area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local historical societies in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, Leadville material, BLM history of northwestern Colorado, and broad Western historical sites. None named a trail meeting all constraints.]
</tool_response>
<think>After searching local historical societies in Denver, Colorado, I found that while there are many urban trails in the area, none of them specifically mention a 19th-century structure. The BLM Cultural Resource Series provides some historical context for the development of northwestern Colorado in the late 19th century, but it does not mention any specific trails. The information provided about Leadville and historical sites in the West does not seem to be directly relevant to the trail in Denver, Colorado that I am looking for. I will try to look for more specific local records or archives that might have information about trails with 19th-century structures in the Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local archives in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, broad Western historical sites, and an Early American Roads and Trails overview. No result matched the joint trail constraints.]
</tool_response>
<think>Despite searching local archives in Denver, Colorado, for trails with 19th-century structures, the search results have not provided a specific trail that matches the given criteria. The information provided about urban trails in Colorado Springs and the historical context of early American roads and trails does not mention a trail near Denver that fits the description. I will try to contact local historical societies or reference facilities at local libraries to get more detailed information about trails with 19th-century structures in the Denver area. These sources might have more specific records or knowledge about such trails.</think>

Stage 2b analysis: new source types, unchanged location

The agent moves from general trail pages to historical records and local sources, but all searches remain centered on Denver. It also fails to include the full set of trail attributes in a single query. As a result, changing the source type neither tests the location assumption nor yields a candidate that satisfies the constraints.

2c. Repeated Denver-specific searches across local sources


```
<tool_call>
{"name": "search", "arguments": {"query": ["contact local historical societies in Denver, Colorado for
trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, a historical-newsletter page, and BLM
history of northwestern Colorado. These offered general context only; no trail matched all constraints.]
</tool_response>
<think>After contacting local historical societies in Denver, Colorado, I have not received any specific
 information about a trail with a 19th-century structure, a length of about 0.50 - 1 mile, a width of 1
- 3 feet, and an elevation gain of 150 - 400 feet. The search results from the historical newsletters
also did not provide any relevant information. The information provided about urban trails in Colorado
Springs and the development of northwestern Colorado in the late 19th century does not mention a trail
that fits the given criteria. It is possible that the trail I am looking for is not well-documented or
might be a local path that is not listed in the general trail databases. I will now try to reach out to
local libraries or historical societies in Denver to see if they have more detailed information or
records about such a trail.</think>
<tool_call>


<tool_response>
[Search results summarized for readability: Returned generic Denver Public Library, Denver Historical
Society, and Denver Open Data Portal pages. None supplied a qualifying trail record.]
</tool_response>
<tool_response>
[Search results summarized for readability: Returned a current-events portal, the West Point Society of
Denver, and an activities calendar. None contained a trail matching the joint constraints.]
</tool_response>
<think>After reaching out to local libraries and historical societies in Denver, I still haven't found
any specific information about a trail with a 19th-century structure, a length of about 0.50 - 1 mile, a
 width of 1 - 3 feet, and an elevation gain of 150 - 400 feet. The information provided about the Denver
 Public Library and Denver Historical Society does not mention any such trail. The calendar of
activities for the West Point Society of Denver also does not provide relevant information. I will now
try to search for more detailed historical records or local community knowledge to find the trail I am
looking for.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["historical records of trails in Denver, Colorado with 19th-
century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, the Old Spanish Trail, and an Early
American Roads and Trails overview. None identified a Denver trail satisfying all attributes.]
</tool_response>
<think>Despite searching for historical records of trails in Denver, Colorado with 19th-century
structures, the search results have not provided a specific trail that matches the given criteria. The
information provided about urban trails in Colorado Springs and the historical context of early American
 roads and trails does not mention a trail near Denver that fits the description. The Old Spanish Trail
and other historical trails are not located in Denver and do not seem to have the specific
characteristics of the trail I am looking for. I will now try to search for more detailed historical
records or local community knowledge to find the trail I am looking for. I will also consider checking
if there are any local historical societies or libraries that have specific information about trails in
the Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["specific historical trails in Denver, Colorado with 19th-
century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, the Old Spanish Trail, Early American
Roads and Trails, and broad Western historical-sites material. No result matched the joint constraints.]
</tool_response>
```

<think>After searching for specific historical trails in Denver, Colorado with 19th-century structures, the search results have not provided a specific trail that matches the given criteria. The information provided about urban trails in Colorado Springs and the historical context of early American roads and trails does not mention a trail near Denver that fits the description. The Old Spanish Trail and other historical trails are not located in Denver and do not seem to have the specific characteristics of the trail I am looking for. Given the lack of relevant information, I will now try to search for more specific local sources or historical records that might mention such a trail. I will also consider checking if there are any local historical societies or libraries that have specific information about trails in the Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local historical records of trails in Denver, Colorado with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Again returned Urban Trails, the Old Spanish Trail, and an Early American Roads and Trails overview. No result matched the full trail description.]
</tool_response>

Stage 2c analysis: rephrasing the same search yields little new evidence

The agent alternates among historical societies, archives, libraries, and forums, but every query retains the Denver constraint. The results therefore repeat the same non-matching material, while the airport distances are never checked again. Malformed or empty tool calls add no evidence and do not alter the search direction.

3a. Community sources, still centered on Denver

<think>
Repeated Denver-specific searches have not produced a matching trail. Since general trail pages and historical records have failed, I will broaden the source type to local community knowledge or forums while retaining the current Denver hypothesis.
</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local community knowledge or forums about trails with 19th-century structure in Denver, Colorado"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, the SITES Community Portal, and a BIPOC
Outdoor Resource Guide. The community-oriented pages did not identify a trail matching all constraints.]
</tool_response>
<think>After searching for local community knowledge or forums about trails with 19th-century structures in Denver, Colorado, the search results have not provided a specific trail that matches the given criteria. The information provided about urban trails in Colorado Springs and the BIPOC Outdoor Resource Guide does not mention a trail near Denver that fits the description. I will continue to search for more specific local sources or historical records that might mention such a trail. I will also consider reaching out to local hiking or outdoor enthusiast groups to see if they have any information about trails with 19th-century structures in the Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local hiking or outdoor enthusiast groups in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned the BIPOC Outdoor Resource Guide, Urban Trails, and BLM history of northwestern Colorado. No qualifying Denver-area trail was identified.]
</tool_response>
<think>The community and outdoor-resource results still do not identify a matching trail. I will search directly for local hiking or outdoor enthusiast groups while retaining the Denver-centered hypothesis.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["contact local hiking or outdoor enthusiast groups in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned the BIPOC Outdoor Resource Guide, Urban Trails, and Trail Anatomy 101. These were general resources and did not identify a trail satisfying all attributes.]
</tool_response>
<think>After contacting local hiking or outdoor enthusiast groups in Denver, Colorado, the search results have not provided a specific trail that matches the given criteria. The information provided

about the BIPOC Outdoor Resource Guide and urban trails in Colorado Springs does not mention a trail near Denver that fits the description. The Trail Anatomy 101 information does not provide any relevant details about trails with 19th-century structures. Given the lack of relevant information, I will continue to search for more specific local sources or historical records that might mention such a trail . I will also consider reaching out to local historical societies or libraries in Denver to see if they have any information about trails with 19th-century structures in the area.</think>

Stage 3a analysis: broader sources do not correct the location

The agent broadens the search to forums, community portals, and local outdoor groups but continues to specify Denver in every query. These additional searches increase the number of tool calls without recovering the Grand Junction clue or combining all trail attributes in one query.

3b. Further archive and community searches

<tool_call>
{"name": "search", "arguments": {"query": ["local historical societies or libraries in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Public Library Book Bikes, Urban Trails, and BLM history of northwestern Colorado. None identified a trail matching the joint constraints.]
</tool_response>
<think>After searching for local historical societies or libraries in Denver, the search results have not provided a specific trail that matches the given criteria. The information provided about public library book bikes and urban trails in Colorado Springs does not mention a trail near Denver that fits the description. The BLM Cultural Resource Series provides some historical context for the development of northwestern Colorado in the late 19th century, but it does not mention any specific trails. Given the lack of relevant information, I will continue to search for more detailed historical records or local community knowledge to find the trail I am looking for. I will also consider reaching out to local historical societies or libraries directly to inquire about trails with 19th-century structures in the Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["contact local historical societies or libraries in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, a historical-newsletter page, and BLM history of northwestern Colorado. No qualifying Denver-area trail was identified.]
</tool_response>
<think>After contacting local historical societies or libraries in Denver, I have not received any specific information about a trail with a 19th-century structure, a length of about 0.50 - 1 mile, a width of 1 - 3 feet, and an elevation gain of 150 - 400 feet. The information provided about urban trails in Colorado Springs and the historical context of the development of northwestern Colorado in the late 19th century does not mention a trail near Denver that fits the description. The search results from the Denver Public Library and Denver Historical Society also do not provide any relevant information. Given the lack of relevant information, I will re-evaluate the plan and consider checking if there are any local community knowledge or forums that might have information about such a trail. I will also consider searching for more detailed historical records or local community knowledge to find the trail I am looking for.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local community knowledge or forums about trails with 19th-century structure in Denver, Colorado"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, the SITES Community Portal, and a BIPOC
Outdoor Resource Guide. None supplied a trail matching all requested attributes.]
</tool_response>
<think>After searching for local community knowledge or forums about trails with 19th-century structures in Denver, Colorado, the search results have not provided a specific trail that matches the given criteria. The information provided about urban trails in Colorado Springs and the BIPOC Outdoor Resource Guide does not mention a trail near Denver that fits the description. The BLM Cultural Resource Series provides some historical context for the development of northwestern Colorado in the late 19th century, but it does not mention any specific trails. Given the lack of relevant information, I will re-evaluate the plan and consider checking if there are any local community knowledge or forums that might have information about such a trail. I will also consider searching for more detailed historical records or local community knowledge to find the trail I am looking for.</think>
<tool_call>

{"name": "search", "arguments": {"query": ["local community knowledge or forums about trails with 19th-century structure in Denver, Colorado"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Repeated Urban Trails, the SITES Community Portal, and the BIPOC Outdoor Resource Guide. No result matched the joint trail constraints.]
</tool_response>
<think>After searching for local community knowledge or forums about trails with 19th-century structures in Denver, Colorado, the search results have still not provided a specific trail that matches the given criteria. The information provided about urban trails in Colorado Springs and the BIPOC Outdoor Resource Guide does not mention a trail near Denver that fits the description. The BLM Cultural Resource Series provides some historical context for the development of northwestern Colorado in the late 19th century, but it does not mention any specific trails. Given the lack of relevant information, I will continue to search for more detailed historical records or local community knowledge to find the trail I am looking for. I will also consider reaching out to local historical societies or libraries directly to inquire about trails with 19th-century structures in the Denver area.</think>
<plan>
1. Contact local historical societies or libraries in Denver, Colorado directly to inquire about trails with 19th-century structures.
2. Search for local community knowledge or forums that might have information about such a trail.
3. Check if there are any historical markers or plaques in the Denver area that might indicate the presence of a trail with a 19th-century structure.
</plan>

<think>
After searching for local community knowledge or forums about trails with 19th-century structures in Denver, Colorado, the search results have not provided a specific trail that matches the given criteria. I will now contact local historical societies or libraries in Denver directly to inquire about trails with 19th-century structures. Additionally, I will search for local community knowledge or forums that might have information about such a trail and check if there are any historical markers or plaques in the Denver area that might indicate the presence of a trail with a 19th-century structure.
</think>
<plan>
1. Contact local historical societies or libraries in Denver, Colorado directly to inquire about trails with 19th-century structures.
2. Search for local community knowledge or forums that might have information about such a trail.
3. Check if there are any historical markers or plaques in the Denver area that might indicate the presence of a trail with a 19th-century structure.
</plan>

<think>
I have contacted local historical societies or libraries in Denver, Colorado, and searched for local community knowledge or forums, but the information provided does not mention a specific trail with a 19th-century structure, a length of about 0.50 - 1 mile, a width of 1 - 3 feet, and an elevation gain of 150 - 400 feet. The search results from the urban trails in Colorado Springs and the BIPOC Outdoor Resource Guide do not mention a trail near Denver that fits the description. The BLM Cultural Resource Series provides some historical context for the development of northwestern Colorado in the late 19th century, but it does not mention any specific trails. I will now check if there are any historical markers or plaques in the Denver area that might indicate the presence of a trail with a 19th-century structure.
</think>

Stage 3b analysis: procedural changes do not update the search

The agent repeats searches across archives, libraries, and community sources and again receives no matching result. Yet it does not treat these failures as evidence against Denver or return to the airport distances. The wording and source type change, but the underlying location assumption remains fixed.

3c. Hiking-group and historical-marker searches

<tool_call>
{"name": "search", "arguments": {"query": ["historical markers or plaques in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, broad Western historical sites, and a History of Pueblo page. None identified a Denver trail matching the full description.]
</tool_response>

<think>After checking for historical markers or plaques in Denver, Colorado, the search results have not provided a specific trail that matches the given criteria. The information provided about urban trails in Colorado Springs and the historical context of the development of northwestern Colorado in the late 19th century does not mention a trail near Denver that fits the description. The Historical Sites That Tell The Stories Of The West and the History of Pueblo do not provide any relevant information about a trail with a 19th-century structure. Given the lack of relevant information, I will continue to search for more detailed historical records or local community knowledge to find the trail I am looking for. I will also consider reaching out to local historical societies or libraries directly to inquire about trails with 19th-century structures in the Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local historical records or community knowledge about trails with 19th-century structure in Denver, Colorado"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, a Columbus/La Raza Park history page, and BLM history of northwestern Colorado. No result matched the joint trail constraints.]
</tool_response>
<think>After searching for local historical records or community knowledge about trails with 19th-century structures in Denver, Colorado, the search results have not provided a specific trail that matches the given criteria. The information provided about urban trails in Colorado Springs and the historical context of the development of northwestern Colorado in the late 19th century does not mention a trail near Denver that fits the description. The Historical Sites That Tell The Stories Of The West and the History of Pueblo do not provide any relevant information about a trail with a 19th-century structure. The information about Columbus Park (La Raza Park) does not mention any trails with the required characteristics. Given the lack of relevant information, I will continue to search for more detailed historical records or local community knowledge to find the trail I am looking for. I will also consider reaching out to local historical societies or libraries directly to inquire about trails with 19th-century structures in the Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local historical societies or libraries in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Public Library Book Bikes, Urban Trails, and BLM history of northwestern Colorado. None contained a qualifying trail record.]
</tool_response>
<think>After searching for local historical societies or libraries in Denver, the search results have not provided a specific trail that matches the given criteria. The information provided about public library book bikes and urban trails in Colorado Springs does not mention a trail near Denver that fits the description. The BLM Cultural Resource Series provides some historical context for the development of northwestern Colorado in the late 19th century, but it does not mention any specific trails. Given the lack of relevant information, I will continue to search for more detailed historical records or local community knowledge to find the trail I am looking for. I will also consider reaching out to local historical societies or libraries directly to inquire about trails with 19th-century structures in the Denver area.</think>

Stage 3c analysis: the search widens around the same location

The agent next consults hiking groups, community sources, and historical-marker pages. These searches remain Denver-specific and produce no evidence for a trail that meets the stated constraints. Changing source types therefore introduces no alternative location.

3d. Final Denver-specific searches

<tool_call>
{"name": "search", "arguments": {"query": ["contact local historical societies or libraries in Denver, Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned Urban Trails, a historical-newsletter page, and BLM history of northwestern Colorado. None identified a trail matching all requested attributes.]
</tool_response>
<think>After contacting local historical societies or libraries in Denver, I have not received any specific information about a trail with a 19th-century structure, a length of about 0.50 - 1 mile, a width of 1 - 3 feet, and an elevation gain of 150 - 400 feet. The information provided about public library book bikes and urban trails in Colorado Springs does not mention a trail near Denver that fits the description. The BLM Cultural Resource Series provides some historical context for the development of northwestern Colorado in the late 19th century, but it does not mention any specific trails. Given the lack of relevant information, I will continue to search for more detailed historical records or

```
local community knowledge to find the trail I am looking for. I will also consider reaching out to local
 hiking or outdoor enthusiast groups to see if they have any information about trails with 19th-century
structures in the Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["local hiking or outdoor enthusiast groups in Denver,
Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Returned the BIPOC Outdoor Resource Guide, Urban Trails, and
BLM history of northwestern Colorado. No qualifying Denver-area trail was identified.]
</tool_response>
<think>After contacting local hiking or outdoor enthusiast groups in Denver, the search results have not
 provided a specific trail that matches the given criteria. The information provided about BIPOC outdoor
 resources and urban trails in Colorado Springs does not mention a trail near Denver that fits the
description. The BLM Cultural Resource Series provides some historical context for the development of
northwestern Colorado in the late 19th century, but it does not mention any specific trails. Given the
lack of relevant information, I will continue to search for more detailed historical records or local
community knowledge to find the trail I am looking for. I will also consider reaching out to local
historical societies or libraries directly to inquire about trails with 19th-century structures in the
Denver area.</think>
<tool_call>
{"name": "search", "arguments": {"query": ["contact local historical societies or libraries in Denver,
Colorado for trails with 19th-century structure"]}}
</tool_call>
<tool_response>
[Search results summarized for readability: Again returned Urban Trails, a historical-newsletter page,
and
BLM history of northwestern Colorado. No result matched the joint trail constraints.]
</tool_response>
```

Stage 3d analysis: the search ends without revisiting its initial assumption

Later searches return the same categories of irrelevant pages. The agent repeatedly notes that no candidate satisfies the constraints but never revisits Grand Junction or re-evaluates its Denver assumption. It thus reaches the end of the search with the original location error intact.

4. Final answer without supporting evidence

```
<think>
After exhausting these Denver-centered sources without finding a match, I will provide the most
plausible candidate under my current location assumption.
</think>
<answer>
Old Elitch's Park Trail
</answer>
```

Stage 4 analysis: the final answer lacks retrieved support

After the Denver-centered searches fail, the agent outputs *Old Elitch's Park Trail*. No preceding result mentions this trail or provides evidence that it matches the requested attributes. The final answer is therefore unsupported and incorrect (EM $= 0$).

Direct comparison: search cost and recovery

BASE. Counting the explicit JSON `search` invocations shown here, BASE makes 25 calls and issues 26 query strings because its first call batches Denver and Chicago. Eight strings are verbatim repeats, and most others only change the source type. All 25 calls retain Denver; none rechecks the airport distances or combines all four trail attributes. Four malformed or empty tool-call turns add no evidence. BASE ultimately returns *Old Elitch's Park Trail* without support (EM $= 0$).
CAFE. CAFE makes two search calls: an initial airport-distance search and one post-feedback query combining Colorado geography with all four trail attributes. The latter retrieves *Old Kiln Trail* near Grand Junction (EM $= 1$). BASE therefore uses $12.5\times$ as many search calls (25 vs. 2) yet fails. In this case, recovery follows from reformulating the query after feedback rather than extending the same loop.