SEARCH-G1: CO-EVOLVING REPRESENTATION-BASED REWARDS FOR SELECTIVE GROUNDED SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Search-augmented language agents should retrieve external information only when necessary and ground their answers in retrieved evidence. Yet terminal correctness assigns equal credit to direct answers from parametric knowledge, evidence-grounded answers, and answers recalled after redundant search. Process supervision and LLM judges provide richer feedback but can require high-cost annotation or inference; low-cost, non-contrastive policy confidence signals do not directly identify whether retrieved evidence changes an answer. We propose **Search-G1**, a reward framework that couples retrieval necessity with evidence reliance through two intervention-calibrated hidden-state readouts for **selective grounded search**. A prompt readout predicts closed-book sufficiency, whose complement proxies policy-relative **retrieval necessity**; an answer-commit readout predicts **evidence reliance** as sensitivity to evidence deletion conditional on the realized reasoning prefix. A correctness-first reward gates evidence-reliance credit by retrieval necessity, rewards sufficient direct answers, and penalizes repeated search. After calibration, scoring needs neither process annotations nor LLM-judge inference. As policy-relative targets, trajectory distributions, and representations may drift, Search-G1 periodically regenerates targets and refits its readouts. Across four QA benchmarks and two scales, Search-G1 achieves competitive task accuracy, with stronger counterfactual evidence dependence and fewer searches in the main comparisons, complemented by independent human audits of answer support.

1 INTRODUCTION

Search agents augment large language models (LLMs) with an external retrieval loop. They issue queries, inspect documents, reason over evidence, and then commit an answer (Shang et al., 2025; Zhan et al., 2026). Recent search-agent systems show that reinforcement learning can induce this behavior (Jin et al., 2025; Song et al., 2025; Gao et al., 2025). Yet the interface only determines what an agent can do. The reward determines which behavior policy optimization reinforces. Terminal correctness cannot distinguish grounded retrieval from redundant search followed by parametric recall, so trajectories that differ sharply in search behavior are reinforced identically. A capable agent should instead perform *selective grounded search*: retrieve when needed, and ground the answer in what was retrieved. This requires two distinct judgments. *Retrieval necessity* (Feng et al., 2026b) asks whether the current policy can answer without retrieval. *Evidence reliance* asks whether a searched answer actually depends on the retrieved context. Rewarding reliance alone can favor redundant retrieval; rewarding necessity alone does not distinguish evidence use from recall after search.

Retrieval need depends on both question structure and policy knowledge (Zhao et al., 2026b). Our initial Qwen2.5-3B-Instruct policy (Ahmed et al., 2025) answers 14% of held-out NQ (Kwiatkowski et al., 2019) questions correctly without retrieval, suggesting scope to avoid redundant search. HotpotQA (Yang et al., 2018) has 79.9% bridge questions, and MuSiQue (Trivedi et al., 2022) contains 2–4-hop questions (Table A3). These structural annotations characterize task demands, not the evidence a particular policy needs. We therefore make both judgments policy-relative: necessity is derived from closed-book sufficiency, while reliance measures answer-stage sensitivity to evidence deletion conditional on the realized reasoning prefix (Section 3.2).

Existing reward designs fall into two broad families, as shown in Figure 1. **External rewards** include sparse answer correctness (Song et al., 2025; Jin et al., 2025), process reward models (Wang et al.,

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

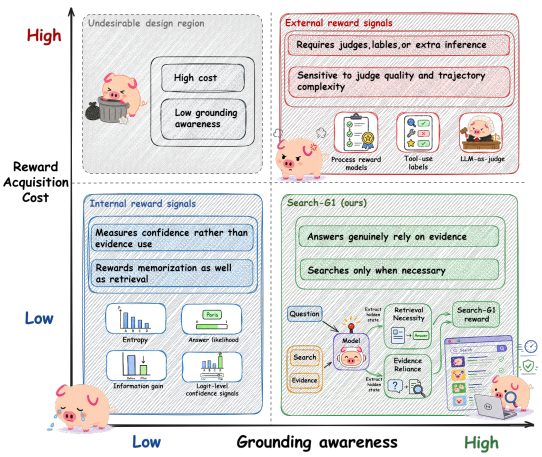


Figure 1: Reward signals for agentic search RL. External rewards range from sparse outcomes to process supervision, tool feedback, and judge signals. Internal rewards provide graded signals but often capture confidence rather than evidence use. These families differ in the trajectory-level credit they assign to correct search behavior: trajectories that are equally correct may warrant different credit depending on whether retrieval was necessary and whether the answer used the retrieved evidence. Search-G1 calibrates two policy readouts through interventions: a prompt readout estimates whether the policy can answer without retrieval, and an answer readout estimates reliance on retrieved evidence.

2025; Xi et al., 2026), tool-grounding feedback (Ma et al., 2025), and LLM-critic or learned-reward feedback (Zhang et al., 2026a; Hu et al., 2025). Richer external signals inspect trajectories more directly, but require annotations or additional model calls and may be brittle on multi-step traces (Liu et al., 2026). **Internal rewards** instead derive graded signals from the policy itself (Du et al., 2019), including entropy (Dong et al., 2026), answer likelihood or information gain (Wang et al., 2026a), and logit-level confidence. Non-contrastive confidence signals do not directly identify evidence use: a model may remain confident because it already knows an answer, even after searching (Mallen et al., 2023; Wang et al., 2023; Tang et al., 2026). Contrastive context-masking rewards assess evidence sensitivity more directly (Tan et al., 2026; Guo et al., 2026). We target a reward that jointly measures retrieval necessity and evidence reliance, with the underlying interventions amortized into readouts that can be refreshed as the policy changes.

To address this gap, we introduce **Search-G1**, a policy-relative reward framework for selective grounded search that combines two intervention-calibrated readouts of the policy's internal representations. A prompt-state readout predicts closed-book sufficiency, whose complement gives retrieval necessity; an answer-commit readout estimates evidence reliance from sensitivity to evidence deletion. Both readouts amortize these interventions during rollout scoring, requiring neither process annotations nor external judge calls at reward time. Their scores form a correctness-first trajectory reward: searched answers receive necessity-conditioned reliance credit, direct answers receive sufficiency credit, and repeated search incurs a bounded cost. This design can restore variation among equally correct trajectories that differ in reliance, branch, or cost. Because reinforcement learning changes policy-relative targets, trajectories, and checkpoint representations, fixed readouts can become stale. Search-G1 therefore periodically regenerates targets and refits both readouts: interventions calibrate, the reward assigns credit, masked GRPO updates the policy, and the next round recalibrates on the latest checkpoint. Across four QA benchmarks and two scales, Search-G1 achieves competitive task accuracy, with stronger counterfactual evidence dependence and fewer searches than other search-enabled baselines in the main grounding evaluations. Independent human-support and fixed-question audits complement these results.

Our main contributions are as follows:

- We formulate selective grounded search as a **policy-relative reward problem**: distinguishing equally correct trajectories requires judging whether retrieval was necessary for the current policy and whether the answer relied on the retrieved evidence, neither of which is available from terminal correctness.
- We propose **Search-G1**, which distills question-level closed-book sufficiency and trajectory-level evidence-deletion sensitivity into internal readouts. A **necessity-gated reward** conditions evidence-reliance credit on closed-book insufficiency while rewarding sufficient direct answers; periodic target regeneration and refitting adapt both readouts to the current policy.
- Experiments show competitive task accuracy, with stronger counterfactual evidence dependence and fewer searches in the main comparisons. Human-support and fixed-question analyses complement these diagnostic results.

2 Related Work

2.1 Search-Agent Rewards and Credit Assignment

Self-RAG uses reflection tokens for retrieval and support judgments (Asai et al., 2024), while Adaptive-RAG routes questions by complexity (Jeong et al., 2024). Search-agent RL instead assigns credit to sampled trajectories. Outcome-driven agents use terminal correctness in Search-R1 (Jin et al., 2025) and longer-horizon training in Beyond Ten Turns (Gao et al., 2025). Denser credit comes from AgentPRM step scores (Xi et al., 2026), GiGPO anchor-state advantages (Feng et al., 2026a), or the CriticSearch retrospective critic (Zhang et al., 2026a); PoU checks citation validity, perturbation sensitivity, and answer–evidence alignment (Ma et al., 2025). Internal shaping uses post-tool entropy (Dong et al., 2026), gold-answer information gain (Wang et al., 2026a), or sibling-context distillation (Liang et al., 2026). Boundary methods use query confidence (Wu et al., 2025), certainty plus retrieval quality (Feng et al., 2026b), disabled/enabled rollout contrasts (Tang et al., 2026), or AKBE's dynamic tool/no-tool probes (Chen et al., 2026). CTRL-RAG contrasts response likelihoods with and without supporting documents (Tan et al., 2026), while concurrent CIPO masks attention to retrieved observations to score subsequent actions (Guo et al., 2026). Concurrent work co-evolves textual guidance (Wang et al., 2026b) or critic supervision (Zhang et al., 2026b). Search-G1 makes current-policy retrieval necessity an explicit gate on evidence-reliance credit, using periodically refreshed hidden-state readouts to amortize retrieval-disabled and evidence-deletion interventions.

2.2 Internal Signals of Evidence Dependence

Internal-state studies examine factual knowledge and its use in context, including relational facts (Petroni et al., 2019), key–value memories (Geva et al., 2021), factual knowledge representations (Hernandez et al., 2023), and knowledge–context reconciliation (Zhao et al., 2026a). Related signals support correctness self-assessment (Kadavath et al., 2022), truthfulness prediction (Azaria & Mitchell, 2023), entity-knowledge estimation (Ferrando et al., 2025), and tool-selection error detection (Healy et al., 2026). Applications include RAG attribution (Qi et al., 2024), hallucination detection (Sun et al., 2025), and factuality decoding (Chuang et al., 2024). Using these signals as training rewards also requires accounting for the policy that produces them. Most of the above methods target fixed-model inference; representation-based exploration extends internal signals to post-training (Tuyls et al., 2026), while PAIR combines a frozen belief-consistency probe with attention correction (Kim et al., 2026). Search-G1 calibrates retrieval-specific targets for closed-book sufficiency and evidence-deletion sensitivity and repeatedly refits both policy-conditioned readouts.

3 Methodology

In this section, we present Search-G1, a reward framework for selective grounded search as shown in Figure 2. We first define the intervention targets and calibrate two representation readouts. Then, we combine their scores into a necessity-gated trajectory reward. Finally, we describe the masked GRPO update, its normalization effects, and periodic readout recalibration.

3.1 Problem Formulation

Given a question q_i, gold answer $y_i^\star$, and search-enabled prompt P_i, the rollout policy samples a group of $G \geq 2$ trajectories,

$$\tau_{i,j} \overset{\text{iid}}{\sim} \pi_{\text{old}}(\cdot \mid P_i), \qquad j = 1, \ldots, G.$$

The rollout policy π_{old} remains fixed while π_θ is updated on the sampled group. Each trajectory contains reasoning, search actions, retrieved observations, and a final answer $\hat{a}_{i,j}$. Let $v_{i,j} \in \{0, 1\}$ indicate validity under the output parser, and let $e_{i,j} = V(\hat{a}_{i,j}, y_i^\star) \in \{0, 1\}$ indicate answer correctness, with $e_{i,j} = 0$ for invalid outputs. The statistic $N_{i,j} \in \mathbb{Z}_{\geq 0}$ counts parsed search actions: $N_{i,j} = 0$ denotes a direct trajectory and $N_{i,j} > 0$ a searched one.

The goal is to convert policy-relative retrieval necessity and evidence sensitivity into correctness-first trajectory credit: favor evidence-sensitive search when needed, direct answering when closed-book knowledge suffices, and discourage repeated search. Because GRPO normalizes within each rollout group, the reward is also designed to differentiate trajectories that a correctness-only signal would tie.

3.2 Intervention-Calibrated Representation Readouts

The two readouts operate at different levels: closed-book sufficiency is shared across a question's rollouts, whereas evidence reliance depends on each searched trajectory. At calibration round m, we

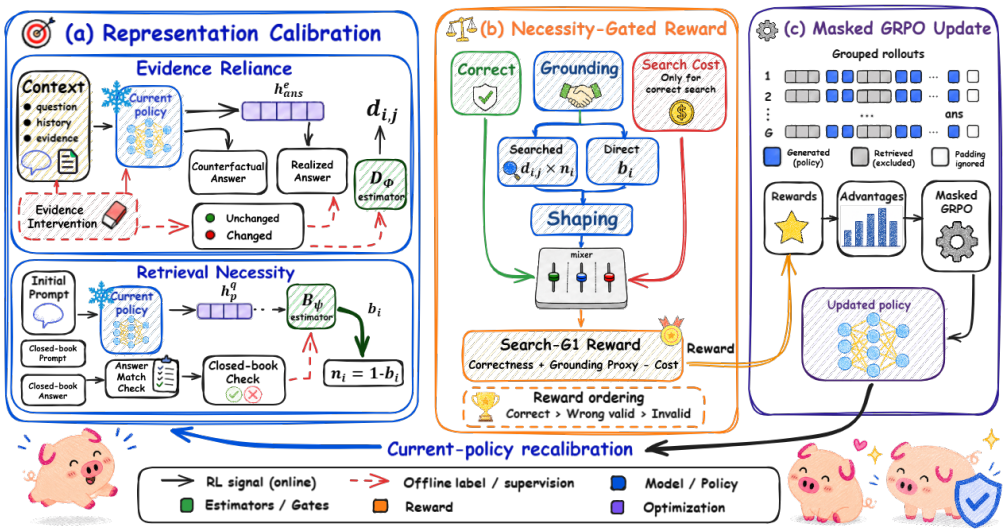


Figure 2: Pipeline of **Search-G1**. We first calibrate two hidden-state readouts from closed-book and evidence-deletion interventions on a frozen policy snapshot. We then combine their scores into necessity-gated trajectory rewards and update the policy with masked GRPO. Finally, we periodically refresh the snapshot, regenerate intervention targets, and refit both readouts as the policy evolves.

freeze the latest checkpoint $\bar{\theta}_m$. It generates counterfactual completions and encodes the prompt-end state $h_{p,i}$ and final-answer state $h_{\mathrm{ans},i,j}$ from original, unperturbed trajectories. Between refits, current trajectories are teacher-forced through the frozen evaluator without gradients; the snapshot and both heads stay fixed, and one evaluator version scores each rollout group.

Evidence Reliance. For a valid searched trajectory, let $C_{i,j}$ be its context immediately before the answer. The intervention $T(C_{i,j})$ removes retrieved-observation blocks while holding the question, search history, and realized reasoning fixed, measuring evidence sensitivity conditional on this prefix. We define

$$\begin{aligned} s^{\mathrm{chg}}_{i,j} &= \mathbf{1}\Big[\nu(\hat{a}_{i,j}) \neq \nu\Big(A^{\mathrm{det}}_{\bar{\theta}_m}(T(C_{i,j}))\Big)\Big], \\ d_{i,j} &= D_{\phi^{(m)}}(h_{\mathrm{ans},i,j}) \in [0,1], \end{aligned} \tag{1}$$

where $A^{\mathrm{det}}_{\bar{\theta}_m}(C)$ is the parsed answer from a greedy continuation of prefix C under the snapshot, and ν applies the task evaluator's deterministic string normalization (lowercasing, article and punctuation removal, and whitespace collapsing). The target compares a sampled answer with a greedy counterfactual; Table A15 audits decoding mismatch. The readout estimates deletion sensitivity and contributes reward only in the correct-searched branch of Equation (6).

Retrieval Necessity. From a fixed retrieval-disabled prompt P^{cb}_i, we obtain a closed-book target and predict it from the actual search-enabled prompt state:

$$\begin{aligned} z_i &= V_{\mathrm{cb}}\Big(A^{\mathrm{det}}_{\bar{\theta}_m}(P^{\mathrm{cb}}_i), y^\star_i\Big) \in \{0,1\}, \\ b_i &= B_{\psi^{(m)}}(h_{p,i}) \in [0,1], \qquad n_i = 1 - b_i. \end{aligned} \tag{2}$$

Here b_i estimates sufficiency under the frozen policy, closed-book prompt, deterministic decoder, and verifier; n_i is therefore policy- and protocol-relative. The target is defined by the retrieval-disabled prompt, while the readout acts on the state that P_i induces before any search. Thus b_i predicts current-policy closed-book sufficiency from the decision state and provides a graded reward-time score without an auxiliary closed-book completion. V_{cb} uses normalized inclusion for free-form completions, which is more permissive than the exact-match criterion used for outcomes; disagreements shift targets toward sufficiency and may bias credit toward direct behavior. Exact conventions are in Section A.3.4.

Let $\mathcal{D}^{(m)}_D$ contain valid searched trajectories and $\mathcal{D}^{(m)}_B$ contain question-level closed-book examples. The two readouts are fitted with

$$\mathcal{L}_{\mathrm{cal}} = \mathbb{E}_{\mathcal{D}^{(m)}_D}\Big[\mathrm{BCE}(d_{i,j}, s^{\mathrm{chg}}_{i,j})\Big] + \mathbb{E}_{\mathcal{D}^{(m)}_B}\big[\mathrm{BCE}(b_i, z_i)\big]. \tag{3}$$

Hard targets are generated on the disjoint calibration pool. The fitted readouts map rollout states to continuous scores and supply both signals in one reward-time forward pass, amortizing a closed-book decode per distinct scored question and an evidence-deleted decode per valid searched trajectory. At each refit, each head selects its layer by validation AUC; fitting, refresh overhead, and label construction are detailed in Sections A.2.1, A.3.1 and A.3.4.

3.3 Necessity-Gated Reward

For a searched trajectory, Search-G1 combines evidence reliance and retrieval necessity into a necessity-gated reliance score

$$g_{i,j} = d_{i,j} n_i. \tag{4}$$

This multiplicative gate is large only when the answer is evidence-sensitive and closed-book sufficiency is low. The product $g_{i,j}$ serves as a graded shaping score for correct searched trajectories. Correct direct trajectories instead use b_i.

Within the correct-searched reward branch, the first search is unpenalized, while any trajectory with two or more search actions activates a unit cost:

$$c_{i,j} = \mathbf{1}[N_{i,j} > 1]. \tag{5}$$

Direct and single-search trajectories have zero cost; repeated search has unit cost regardless of further calls. Under the four-round cap, this bounded term distinguishes single from repeated retrieval, preserving the correct-searched lower bound of $1 - \eta$ and the correctness-first reward tiers.

Let r_{inv} and r_{wrong} denote the rewards for invalid and wrong outputs. The Search-G1 reward is

$$r_{i,j} = \begin{cases} r_{\text{inv}}, & v_{i,j} = 0, \\ r_{\text{wrong}}, & v_{i,j} = 1,\ e_{i,j} = 0, \\ 1 + \lambda_g g_{i,j} - \eta c_{i,j}, & v_{i,j} = 1,\ e_{i,j} = 1,\ N_{i,j} > 0, \\ 1 + \alpha_{\text{cb}} b_i, & v_{i,j} = 1,\ e_{i,j} = 1,\ N_{i,j} = 0. \end{cases} \tag{6}$$

For a correct searched trajectory and a correct direct trajectory from the same question, the searched trajectory ranks higher before—and therefore after—group normalization exactly when

$$\lambda_g d_{i,j}(1 - b_i) - \eta c_{i,j} > \alpha_{\text{cb}} b_i. \tag{7}$$

Higher closed-book sufficiency thus shifts credit toward direct answering, while reliance and necessity must jointly offset search cost; refitting the sufficiency readout against the latest checkpoint keeps this boundary tied to current capability. The tiers themselves hold whenever $r_{\text{inv}} < r_{\text{wrong}} < 1 - \eta$, $\eta \in [0, 1)$, and $\lambda_g, \alpha_{\text{cb}} \geq 0$: every correct valid trajectory outranks every wrong one, which outranks every invalid output. Nonnegativity is all this requires; our coefficient values are listed in Section A.2.1.

3.4 Group-Relative Credit under Masked GRPO

Under masked GRPO, each trajectory receives the group-relative advantage (Shao et al., 2024),

$$\hat{A}_{i,j} = \frac{r_{i,j} - \mu_i}{\sigma_i + \delta}, \tag{8}$$

where μ_i and σ_i are the group mean and sample standard deviation and $\delta > 0$ stabilizes normalization.

We use the standard clipped GRPO update with Search-G1 rewards. Let $M_{i,j,t} = 1$ for reasoning, search-action, and final-answer tokens and 0 for prompts, retrieved observations, and padding. We write $\langle \cdot \rangle_{\mathcal{M},\mathcal{B}}$ for the masked token average over batch $\mathcal{B}$ and maximize

$$\begin{aligned} \ell^{\text{clip}}_{i,j,t} &= \min\Big\{ \rho_{i,j,t} \hat{A}_{i,j}, \text{clip}(\rho_{i,j,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,j} \Big\}, \\ \mathcal{J}_{\text{G1}}(\theta) &= \Big\langle \ell^{\text{clip}}_{i,j,t} + \alpha_{\text{H}} \mathcal{H}(\pi_\theta(\cdot \mid \tau_{i,j,<t})) - \beta_{\text{KL}} \widehat{\text{KL}}^{\text{LV}}_{i,j,t} \Big\rangle_{\mathcal{M},\mathcal{B}}, \end{aligned} \tag{9}$$

Here $\rho_{i,j,t}$ is the current-to-rollout-policy token likelihood ratio, ϵ is the clipping width, and $\alpha_{\text{H}}, \beta_{\text{KL}}$ weight entropy and reference-policy KL regularization. The reward enters through $\hat{A}_{i,j}$, so its within-group distinctions determine the credit assigned to policy-generated tokens. The ratio, KL estimator, and distributed averaging rule are specified in Section A.1.

Effect of GRPO Normalization. Centering removes group-wide additive terms, so the group-constant n_i and b_i act differently from the trajectory-varying $d_{i,j}$: a multiplicative gate can still reweight searched credit against the other branches. To characterize this, let $\gamma \in [0,1]$ replace n_i in the correct-searched reliance term only, so that $\gamma = n_i$ recovers Search-G1 and $\gamma = 1$ removes the gate; let $\mathbf{u}_i$ and $\mathbf{v}_i$ be the group-centered versions of the $\gamma = 0$ reward vector and of $x_{i,j} = d_{i,j}\mathbf{1}[v_{i,j} = e_{i,j} = 1,\ N_{i,j} > 0]$, and set $\mathbf{z}_i(\gamma) = \mathbf{u}_i + \lambda_g \gamma \mathbf{v}_i$.

Proposition 1 (Normalization property of the necessity gate)**.** *Assume $\lambda_g > 0$. For any $\gamma \in [0,1]$ with $\mathbf{z}_i(\gamma) \neq \mathbf{0}$ and any $\delta \geq 0$, group normalization yields*

$$\mathbf{A}_i^{\delta}(\gamma) = \frac{\mathbf{z}_i(\gamma)}{\|\mathbf{z}_i(\gamma)\|_2/\sqrt{G-1} + \delta}, \tag{10}$$

which preserves the direction of $\mathbf{z}_i(\gamma)$; when $\delta = 0$, normalization also removes its scale. Consequently: (i) *in an all-valid, all-correct, all-searched, equal-cost group with $\mathbf{v}_i \neq \mathbf{0}$, $\mathbf{u}_i = \mathbf{0}$, so every $\gamma > 0$ gives the same rollout ranking, advantage signs, and credit direction, determined by $d_{i,j}$;* (ii) *for distinct γ_1, γ_2 with $\mathbf{z}_i(\gamma_1), \mathbf{z}_i(\gamma_2) \neq \mathbf{0}$, the normalized credit directions differ whenever $\mathbf{u}_i$ and $\mathbf{v}_i$ are non-collinear.*

This division of labor is what Equation (6) implements: $d_{i,j}$ and $c_{i,j}$ separate trajectories inside the correct-searched branch, while n_i and its complement b_i act across branches, making the direct–search comparison monotone in closed-book sufficiency whenever both behaviors occur (Equation (7)). If $\mathbf{z}_i(\gamma) = \mathbf{0}$, the centered reward is constant and yields no within-group advantage; Section A.1 gives the proof, exact sign-change criterion, and boundary cases.

Current-Policy Recalibration. At each round, fixed calibration IDs are rerun under $\bar{\theta}_m$ to regenerate trajectories, targets, and states before both heads are refit and installed atomically. Here π_{old} is batch-local, whereas $\bar{\theta}_m$ persists across updates. Each refresh updates the intervention targets, their hidden-state representations, and the fitted heads, adapting both measurements to the latest policy. Schedules and overhead are given in Section A.2.1.

4 Evaluation

We evaluate the overall accuracy–grounding–search trade-off, then examine whether the readouts capture their intended targets and provide useful credit as the policy evolves. Component ablations test the roles of the reward terms and evaluator refresh.

4.1 Experimental Setup

Models and retrieval environment. We evaluate Qwen2.5-3B/7B-Instruct (Ahmed et al., 2025) in a fixed Wikipedia-2018 retrieval environment with `e5-base-v2`, FAISS, three passages per retrieval, and at most four rounds. Training/calibration, validation/architecture-selection, and final-evaluation IDs are disjoint; validation selects reward and refresh settings, and checkpoints are fixed before held-out greedy evaluation without test-based selection, as detailed in Section A.2.1.

Benchmarks and metrics. We report exact match (**EM**↑) on NQ (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). On NQ and 2Wiki, grounding evaluation replaces an answer-bearing fact for originally correct searched trajectories. Searched-only **D-Inv**↓ is the fraction preserving the original answer; **TC**↑ combines correct-direct retention with graded searched-answer response; and **Search/Q**↓ counts generated search markers per question. These diagnostics are not used as rollout rewards or readout-fitting targets; their validation-split values are used only for configuration selection. These diagnostics measure counterfactual evidence dependence. Because TC and D-Inv condition on policy outcomes, we complement them with searched-only accounting on fixed IDs and an unconditional common-question decomposition (Tables A14 and A18), alongside surface controls and independent human support. Tables 1 and 2 report NQ/2Wiki results; Section A.3.1 defines the metrics.

Baselines. We compare with two non-trained references: **Base** (Ahmed et al., 2025), the closed-book model without retrieval, and **Prompted Search** (Jin et al., 2025), which uses the same retrieval interface without policy updates. The training baselines cover complementary credit mechanisms: **Search-R1** (Jin et al., 2025) optimizes terminal correctness with a direct-answer bonus in our port;

Table 1: Comparison of Search-G1 and baseline methods across task accuracy, grounding, and search cost at two model scales. EM columns are dataset-specific; TC, D-Inv, and Search/Q are measured on NQ. Base/Prompted Search are deterministic. Arrows show deltas from Prompted Search; TC/D-Inv are n/a for Base.

Method	Signal	Task Utility				Grounding (NQ)		Search Cost (NQ)
		NQ (↑)	HotpotQA (↑)	2WikiMultiHopQA (↑)	MuSiQue (↑)	TC (↑)	D-Inv (↓)	Search/Q (↓)
Qwen2.5-3B-Instruct								
Base	None	0.142	0.153	0.212	0.029	n/a	n/a	0.000
Prompted Search	Retrieval	0.250	0.199	0.238	0.057	0.930	0.069	2.215
Search-R1 (Jin et al., 2025)	Outcome+Direct	0.352 ↑0.102	0.248 ↑0.049	0.267 ↑0.029	0.087 ↑0.030	0.917 ↓0.013	0.049 ↓0.020	3.281 ↑1.066
GiGPO (Feng et al., 2026a)	State	0.422 ↑0.172	0.296 ↑0.097	0.330 ↑0.092	0.126 ↑0.069	0.935 ↑0.005	0.047 ↓0.022	3.131 ↑0.916
AgentPRM (Xi et al., 2026)	Process	0.380 ↑0.130	0.293 ↑0.094	0.328 ↑0.090	0.112 ↑0.055	0.958 ↑0.028	0.042 ↓0.027	3.340 ↑1.125
GPT-5 Judge (Singh et al., 2025)	External Judge	0.402 ↑0.152	0.280 ↑0.081	0.324 ↑0.086	0.101 ↑0.044	0.946 ↑0.016	0.038 ↓0.031	3.238 ↑1.023
ARPO (Dong et al., 2026)	Entropy	0.389 ↑0.139	0.273 ↑0.074	0.304 ↑0.066	0.109 ↑0.052	0.924 ↓0.006	0.055 ↓0.014	2.679 ↑0.464
IGPO (Wang et al., 2026a)	Information Gain	0.414 ↑0.164	0.292 ↑0.093	0.355 ↑0.117	0.137 ↑0.080	0.934 ↑0.004	0.056 ↓0.013	3.028 ↑0.813
Search-E1 (Liang et al., 2026)	Self-distillation	0.437 ↑0.187	**0.314** ↑0.115	**0.372** ↑0.134	0.142 ↑0.085	0.939 ↑0.009	0.054 ↓0.015	2.980 ↑0.765
SAAS (Tang et al., 2026)	Rollout Boundary	0.428 ↑0.178	0.301 ↑0.102	0.355 ↑0.117	0.138 ↑0.081	0.951 ↑0.021	0.045 ↓0.024	2.122 ↓0.093
β-GRPO (Wu et al., 2025)	Query Confidence	0.419 ↑0.169	0.295 ↑0.096	0.344 ↑0.106	0.129 ↑0.072	0.948 ↑0.018	0.047 ↓0.022	2.227 ↑0.012
KbPO (Feng et al., 2026b)	Certainty+Quality	0.431 ↑0.181	0.305 ↑0.106	0.358 ↑0.120	0.140 ↑0.083	0.955 ↑0.025	0.043 ↓0.026	2.160 ↓0.055
Search-G1 (ours)	Representation	**0.453** ↑0.203	0.306 ↑0.107	0.368 ↑0.130	**0.143** ↑0.086	**0.976** ↑0.046	**0.034** ↓0.035	**2.094** ↓0.121
Qwen2.5-7B-Instruct								
Base	None	0.212	0.203	0.250	0.041	n/a	n/a	0.000
Prompted Search	Retrieval	0.352	0.292	0.274	0.074	0.936	0.061	2.171
Search-R1 (Jin et al., 2025)	Outcome+Direct	0.417 ↑0.065	0.382 ↑0.090	0.422 ↑0.148	0.154 ↑0.080	0.932 ↓0.004	0.046 ↓0.015	3.090 ↑0.919
GiGPO (Feng et al., 2026a)	State	0.464 ↑0.112	0.416 ↑0.124	0.436 ↑0.162	0.189 ↑0.115	0.947 ↑0.011	0.043 ↓0.018	2.930 ↑0.759
AgentPRM (Xi et al., 2026)	Process	0.455 ↑0.103	0.405 ↑0.113	0.430 ↑0.156	0.205 ↑0.131	0.963 ↑0.027	0.037 ↓0.024	3.080 ↑0.909
GPT-5 Judge (Singh et al., 2025)	External Judge	0.470 ↑0.118	0.410 ↑0.118	0.440 ↑0.166	0.215 ↑0.141	0.952 ↑0.016	0.033 ↓0.028	2.980 ↑0.809
ARPO (Dong et al., 2026)	Entropy	0.462 ↑0.110	0.398 ↑0.106	0.420 ↑0.146	0.208 ↑0.134	0.931 ↓0.005	0.050 ↓0.011	2.430 ↑0.259
IGPO (Wang et al., 2026a)	Information Gain	0.480 ↑0.128	0.425 ↑0.133	0.455 ↑0.181	0.235 ↑0.161	0.942 ↑0.006	0.050 ↓0.011	2.760 ↑0.589
Search-E1 (Liang et al., 2026)	Self-distillation	0.495 ↑0.143	0.440 ↑0.148	0.465 ↑0.191	0.245 ↑0.171	0.947 ↑0.011	0.048 ↓0.013	2.720 ↑0.549
SAAS (Tang et al., 2026)	Rollout Boundary	0.486 ↑0.134	0.432 ↑0.140	0.458 ↑0.184	0.239 ↑0.165	0.958 ↑0.022	0.038 ↓0.023	2.050 ↓0.121
β-GRPO (Wu et al., 2025)	Query Confidence	0.478 ↑0.126	0.424 ↑0.132	0.448 ↑0.174	0.231 ↑0.157	0.953 ↑0.017	0.040 ↓0.021	2.164 ↓0.007
KbPO (Feng et al., 2026b)	Certainty+Quality	0.490 ↑0.138	0.436 ↑0.144	0.462 ↑0.188	0.242 ↑0.168	0.962 ↑0.026	0.036 ↓0.025	2.106 ↓0.065
Search-G1 (ours)	Representation	**0.512** ↑0.160	**0.450** ↑0.158	**0.475** ↑0.201	**0.254** ↑0.180	**0.981** ↑0.045	**0.028** ↓0.033	**1.860** ↓0.311

GiGPO (Feng et al., 2026a) adds anchor-state groups for stepwise relative credit; **AgentPRM** (Xi et al., 2026) uses a fixed external process-reward model; our **GPT-5 Judge** (Singh et al., 2025) scores completed trajectories with a fixed bespoke evidence-support rubric; **ARPO** (Dong et al., 2026) uses post-tool entropy for adaptive rollouts and attribution; **IGPO** (Wang et al., 2026a) rewards turn-level gains in gold-answer probability; and **Search-E1** (Liang et al., 2026) combines GRPO with offline self-distillation from sibling contexts. We additionally compare with three boundary-oriented methods: **SAAS** (Tang et al., 2026) contrasts search-disabled and search-enabled rollouts to learn when to search; **β-GRPO** (Wu et al., 2025) gates correctness rewards by minimum query-token confidence; and **KbPO** (Feng et al., 2026b) combines closed-book certainty and retrieval quality into behavior rewards. All actors share the model, retriever, and interaction/update cap; details and auxiliary costs are in Section A.2.1 and Table A2.

Implementation details. Trained methods use four independent seeds; the main comparisons, ablations, and sensitivity studies report means across these runs. Fixed-rollout methods sample groups of five trajectories for at most four rounds; adaptive baselines reallocate samples within the same cap. GRPO uses batch size 128, learning rate 10^{-6}, and one clipped-update epoch ($\epsilon = 0.2$), with actor loss restricted to policy-generated tokens. Every 25 updates, both readouts are asynchronously refit on a disjoint 300-question pool, selecting architecture and layer by validation AUC. Reward/refresh settings and final checkpoints are fixed on validation before greedy held-out evaluation. Experiments use eight NVIDIA H20 GPUs (96 GB) and eight A100 GPUs (80 GB), with complete settings in Section A.2.1.

4.2 Main Results

Search-G1 achieves competitive task accuracy, with higher TC, lower D-Inv, and fewer searches than other search-enabled baselines in the grounding comparisons (Tables 1 and 2).

Task accuracy across benchmarks. As shown in Table 1, Search-G1 improves mean EM over Prompted Search on every benchmark and at both scales, by 0.086–0.203 at 3B and 0.158–0.201 at 7B. It has the highest 7B mean on all four benchmarks; at 3B, it leads NQ and MuSiQue and remains within $0.008/0.004$ of the best HotpotQA/2Wiki mean. This pattern spans single-hop and multi-hop settings rather than concentrating on one benchmark.

Grounding and search economy. The NQ results in Table 1 show the targeted joint behavior: at both scales, Search-G1 has the highest mean EM and TC, the lowest mean D-Inv, and the lowest mean Search/Q among search-enabled methods. Relative to Prompted Search, it raises EM by $0.203/0.160$ while reducing Search/Q by $0.121/0.311$. On 2Wiki, Table 2 shows the same direction: relative to Search-R1, Search-G1 raises mean EM by $0.101/0.053$ and TC by $0.059/0.057$, lowers D-Inv by $0.027/0.024$, and uses $0.590/0.703$ fewer searches at 3B/7B.

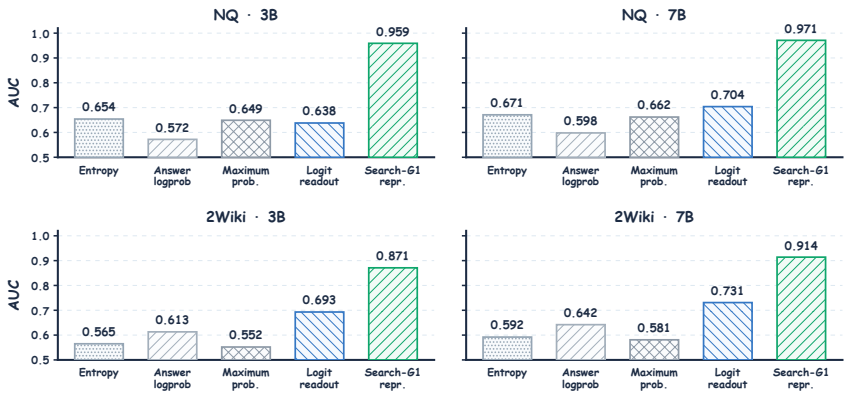


Figure 3: Comparison of evidence-reliance predictors AUC on NQ and 2WikiMultiHopQA.

Table 2: Comparison of accuracy, grounding, and search cost on 2WikiMultiHopQA. Full-split greedy results.

Method	EM↑	TC↑	D-Inv↓	Search/Q↓
Qwen2.5-3B-Instruct				
Prompted Search	0.238	0.877	0.085	2.951
Search-R1	0.267	0.913	0.058	3.402
Search-G1	**0.368**	**0.972**	**0.031**	**2.812**
Qwen2.5-7B-Instruct				
Prompted Search	0.274	0.887	0.079	2.892
Search-R1	0.422	0.921	0.050	3.198
Search-G1	**0.475**	**0.978**	**0.026**	**2.495**

Table 3: Ablation study of reward components and readout refresh on 2WikiMultiHopQA. Entries report EM/TC/Search/Q.

Variant	Qwen2.5-3B	Qwen2.5-7B
Full Search-G1	0.368 / 0.972 / 2.812	0.475 / 0.978 / 2.495
No-refresh evaluator	0.357 / 0.906 / 3.225	0.466 / 0.922 / 2.846
w/o search cost	0.338 / 0.920 / 3.576	0.449 / 0.934 / 3.144
w/o reliance shaping	0.354 / 0.797 / 1.992	0.463 / 0.829 / 1.798
w/o searched gate	0.347 / 0.877 / 3.119	0.444 / 0.897 / 2.756
w/o direct bonus	0.352 / 0.915 / 2.951	0.451 / 0.936 / 2.612
w/o both necessity terms	0.339 / 0.861 / 3.286	0.436 / 0.881 / 2.905
Shuffled searched gate	0.341 / 0.869 / 3.204	0.439 / 0.889 / 2.838

Hierarchical inference. Across four independent training seeds, hierarchical bootstrap with Holm correction supports improvements over Search-R1 on all four NQ endpoints; EM intervals against the post-hoc metric-wise baseline envelope include zero (Table A1). These results support competitive accuracy alongside the grounding and search-efficiency gains; Section A.2.1 details the statistical protocol.

4.3 Analysis

We examine readout fidelity, the alignment of group-relative credit with evidence support, and reliability across policy updates. Behavioral and independent human audits assess the resulting search use and answer support.

Readout and label validity. Figure 3 shows answer-commit AUC $0.959/0.971$ on NQ and $0.871/0.914$ on 2Wiki. Additionally removing retained reasoning or matching decoding yields 93–96% agreement with defaults and positive-rate shifts at most 0.03 (Tables A6 and A15).

Surface controls. In the 3B audit, answer-commit AUC remains $0.934/0.842$ after matching on search count and response length; shuffled labels yield near-chance AUC (Table A16).

Credit beyond terminal correctness. Figure 4 reports nonzero variation in 46.6–48.7% of same-EM NQ groups and 91.8–92.4% on 2Wiki; in the 3B/2Wiki control, advantage–support correlation is 0.41 versus 0.02 after permutation (Table A19). The permutation control tests credit–support association beyond the presence of reward variation.

Recalibration across policy updates. Readouts must remain informative across updates. At checkpoint 150, frozen reliance and necessity AUCs at 3B/7B fall to $0.589/0.608$ and $0.482/0.501$; refitting retains $0.878/0.891$ and $0.786/0.804$ on 2WikiMultiHopQA (Figure 5).

Policy boundary and realized cost. On its closed-book-correct slices, Search-G1 has repeated-search rates of $0.040/0.028$. On fixed Base-CB-wrong questions, valid-conditioned EM reaches $0.327/0.356$ with fewer searches than Search-R1 (Tables A7 and A8). On NQ, policy-generated and total response-side tokens fall 43–44% and 16–19% versus Search-R1 (Tables A4 and A5).

Independent grounding. Human judgments use neither perturbations nor readouts: Search-G1 reaches 0.842 Support@Correct on independent correct-searched pools (Table A10). On 600

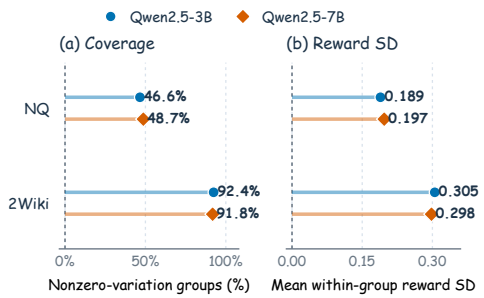


Figure 4: Reward variation within same-EM groups. Left: coverage of nonzero-variance groups; right: their mean within-group reward SD.

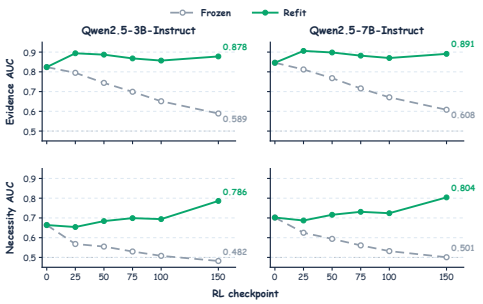


Figure 5: Evolution of readout AUC across policy checkpoints on 2WikiMultiHopQA. Top: evidence reliance; bottom: necessity; left: 3B; right: 7B.

shared NQ IDs, its unconditional grounded accuracy is $0.360\,[0.322, 0.399]$, comparable to GPT-5 Judge at $0.355\,[0.317, 0.394]$, with the same rounded unsupported rate (Table A18). This common-denominator check complements conditional support; Table A9 reports transfer results.

Together, these analyses connect accurate intervention readouts to credit aligned with human-assessed support, while showing that recalibration preserves readout reliability as the policy evolves.

4.4 Ablations

Component ablations. In the four-run means, removing search cost increases Search/Q and lowers EM: more retrieval need not improve accuracy. Removing reliance shaping reduces search but incurs the largest TC loss (Table 3), while removing refresh lowers EM/TC and increases search.

Necessity factorization. As Proposition 1 predicts, n_i can change credit direction in heterogeneous groups, but not in all-correct, equal-cost searched groups. Removing the gate lowers EM/TC and raises Search/Q; shuffling n_i also degrades performance (Section A.1).

Coefficient and refresh sensitivity. Tables A12 and A13 show how reward weights and refresh frequency balance objectives: $\lambda_g = 0.8$ raises EM at lower TC and more search, $\eta = 0.05$ increases Search/Q, and $K = 50$ reduces Search/Q but lowers EM and TC.

Unified accounting. On $1{,}024$ shared IDs, Search-G1 leads EM (0.355), searched-only TC (0.655), and EM per marker (0.225) at 1.580 searches per question. Shared IDs align questions, and S-TC excludes direct answers; correct-searched subsets remain policy-dependent (Table A14).

4.5 Case Study

Figure A2 shows three selected 2Wiki entity, bridge, and compositional cases. Search-G1 completes each relation chain with one search; comparison trajectories miss a relation or fail to use retrieved evidence. These cases illustrate relation-chain completion and complement reward diagnostics.

5 Conclusion

In this paper, we present Search-G1, which conditions evidence-reliance credit on policy-relative retrieval necessity. Two intervention-calibrated readouts amortize estimates of closed-book sufficiency and answer-stage evidence-deletion sensitivity, supporting a correctness-first reward that favors correct direct answers when knowledge suffices, rewards evidence-sensitive search when needed, and penalizes repeated retrieval. Periodic target regeneration and refitting adapt both signals as the policy evolves, without external judge calls during policy optimization. Across four QA benchmarks and two scales, Search-G1 achieves competitive task accuracy, stronger counterfactual evidence dependence, and fewer searches in the main comparisons. Ablations and independent human-support audits further support the reward design in the evaluated QA settings.

Limitations. Our evaluation covers short-form QA with static Wikipedia retrieval and two Qwen scales; training uses verifiable gold answers. Open-ended, live-web, multilingual, multimodal, larger-scale, and alternative-RL settings are natural extensions. Efficiency is measured through search and token use. Reliance captures prefix-conditional deletion sensitivity, which may include distribution-shift effects.

AI USE STATEMENT

Generative AI tools assisted language editing and provided critical feedback on methodological exposition, mathematical and logical consistency, claim scope, and interpretation wording. This authoring assistance did not generate experimental data or reported numerical results. Separately, the GPT-5 Judge baseline uses model inference to provide reward feedback, as specified in the experimental setup. The authors reviewed all AI-suggested revisions, made the final scientific judgments, and take full responsibility for the paper.

ETHICS STATEMENT

This work uses publicly available models, benchmarks, and retrieval corpora. Human annotators assessed only retrieved evidence and model outputs and were not asked to provide personal or sensitive information. Search agents may inherit errors and biases from their base models, retrievers, and corpora; therefore, our grounding metrics should not be interpreted as guarantees of factuality or safety, particularly in high-stakes applications.

REPRODUCIBILITY STATEMENT

The method and objective are specified in Section 3 and Section A.1. Data, retrieval, optimization, reward, parser, and readout-refresh settings are documented in Section A.2.1; Section A.3.1 specifies the statistical analyses. An anonymized code artifact is included with the supplementary material and provides the implementation, configurations, and evaluation scripts used for the reported experiments.

REFERENCES

Imtiaz Ahmed, Sadman Islam, Partha Protim Datta, Imran Kabir, Md Naseef Ur Rahman Chowdhury, and Ahshanul Haque. Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potentioal to surpass all competitors. 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Dingwei Chen, Zefang Zong, Zhipeng Ma, Leo Luo, Yang Li, Chengming Li, Peng Chen, and Jie Jiang. Efficient agentic reinforcement learning with on-policy intrinsic knowledge boundary enhancement. *arXiv preprint arXiv:2605.26952*, 2026.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 16981–17017, 2026.

Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

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

Tao Feng, Xinke Jiang, Xinyan Hu, Yonggang Zhang, Zhen Tao, Wentao Zhang, Boyang Liu, Wenhao Jiang, and Chao Wu. Trust within? seek beyond? knowledge boundary aware policy optimization for agentic search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27664–27682, 2026b.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations*, volume 2025, pp. 77965–78000, 2025.

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

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 5484–5495, 2021.

Xingyu Guo, Wei Chen, Linlin Yang, and Baochang Zhang. Contextual information policy optimization for search agents. *arXiv preprint arXiv:2608.06128*, 2026.

Kait Healy, Bharathi Srinivasan, Visakh Madathil, and Jing Wu. Internal representations as indicators of hallucinations in agent tool selection. *arXiv preprint arXiv:2601.05214*, 2026.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Ziyou Hu, Zhengliang Shi, Minghang Zhu, Haitao Li, Teng Sun, Pengjie Ren, Suzan Verberne, and Zhaochun Ren. Openreward: Learning to reward long-form agentic tasks via reinforcement learning. *arXiv preprint arXiv:2510.24636*, 2025.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 7036–7050, 2024.

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. *arXiv preprint arXiv:2503.09516*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Wonjoong Kim, Yeonjun In, Sangwu Park, Dongha Lee, and Chanyoung Park. Pair: Prefix-aware internal reward model for multi-turn agent optimization. *arXiv preprint arXiv:2605.17877*, 2026.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Zihan Liang, Yufei Ma, Ben Chen, Zhipeng Qian, Xuxin Zhang, Huangyu Dai, and Lingtao Mao. Search-e1: Self-distillation drives self-evolution in search-augmented reasoning. *arXiv preprint arXiv:2605.22511*, 2026.

Xuannan Liu, Xiao Yang, Zekun Li, Peipei Li, and Ran He. Agenthallu: Benchmarking automated hallucination attribution of llm-based agents. *arXiv preprint arXiv:2601.06818*, 2026.

SHengjie Ma, Chenlong Deng, Jiaxin Mao, Jiadeng Huang, Teng Wang, Junjie Wu, Changwang Zhang, et al. Pou: Proof-of-use to counter tool-call hacking in deepresearch agents. *arXiv e-prints*, pp. arXiv–2510, 2025.

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 *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6037–6053, 2024.

Yu Shang, Yu Li, Keyu Zhao, Likai Ma, Jiahe Liu, Fengli Xu, and Yong Li. Agentsquare: Automatic llm agent search in modular design space. In *International Conference on Learning Representations*, volume 2025, pp. 3841–3865, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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. *arXiv preprint arXiv:2503.05592*, 2025.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, volume 2025, pp. 50250–50279, 2025.

Zhehao Tan, Yihan Jiao, Dan Yang, Junjie Wang, Duolin Sun, Jie Feng, Xidong Wang, Lei Liu, Yue Shen, Jian Wang, et al. Ctrl-rag: Contrastive likelihood reward based reinforcement learning for context-faithful rag models. *arXiv preprint arXiv:2603.04406*, 2026.

Yunbo Tang, Chengyi Yang, Shiyu Liu, Zhishang Xiang, Zerui Chen, Qinggang Zhang, and Jinsong Su. Saas: Self-aware reinforcement learning for over-search mitigation in agentic search. *arXiv preprint arXiv:2605.29796*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Jens Tuyls, Dylan Foster, Akshay Krishnamurthy, and Jordan Ash. Representation-based exploration for language models: From test-time to post-training. In *International Conference on Learning Representations*, volume 2026, pp. 120621–120645, 2026.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*, 2023.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, et al. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations*, volume 2026, pp. 104502–104529, 2026a.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*, 2025.

Xinyu Wang, Hanwei Wu, Jingwei Song, Shuyuan Zhang, Jiayi Zhang, Fanqi Kong, Tung Sum Thomas Kwok, Xiao-Wen Chang, Yuyu Luo, Chenglin Wu, et al. Co-evolution of policy and internal reward for language agents. *arXiv preprint arXiv:2604.03098*, 2026b.

Peilin Wu, Mian Zhang, Xinlu Zhang, Xinya Du, and Zhiyu Chen. Search wisely: Mitigating sub-optimal agentic searches by reducing uncertainty. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19734–19745, 2025.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, et al. Agentprm: Process reward models for llm agents via step-wise promise and progress. In *Proceedings of the ACM Web Conference 2026*, pp. 4184–4195, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Qiusi Zhan, Angeline Budiman-Chan, Abdelrahman Zayed, Xingzhi Guo, Daniel Kang, and Joo-Kyung Kim. Safesearch: Do not trade safety for utility in llm search agents. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2800–2815, 2026.

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 *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12272–12290, 2026a.

Yudi Zhang, Meng Fang, Zhenfang Chen, and Mykola Pechenizkiy. Self-evolving llm agents with in-distribution optimization. *arXiv preprint arXiv:2606.07367*, 2026b.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. *Advances in Neural Information Processing Systems*, 38:102978–103012, 2026a.

Linxi Zhao, Sofian Zalouk, Christian Belardi, Justin Lovelace, Jin Zhou, Ryan Noonan, Dongyoung Go, Kilian Weinberger, Yoav Artzi, and Jennifer Sun. Pre-training limited memory language models with internal and external knowledge. In *International Conference on Learning Representations*, volume 2026, pp. 60117–60152, 2026b.

A Appendix

A.1 Methodology supplements

Token likelihood ratio. For a selected token $w_{i,j,t}$, the current-to-rollout-policy likelihood ratio in Equation (9) is

$$\rho_{i,j,t} = \frac{\pi_\theta(w_{i,j,t} \mid \tau_{i,j,<t})}{\pi_{\text{old}}(w_{i,j,t} \mid \tau_{i,j,<t})}. \tag{11}$$

Distributed implementation of the masked objective. For the masked average in Equation (9), let $\mathcal{B}_{s,w}$ denote the rank-local micro-batch on data-parallel worker w at gradient-accumulation step s. With W workers and S accumulation steps, and assuming every local micro-batch contains at least one selected token, the implementation-level reduction is

$$\langle f_{i,j,t} \rangle_{\mathcal{M},\mathcal{B}} = \frac{1}{SW} \sum_{s=1}^{S} \sum_{w=1}^{W} \frac{\sum_{(i,j)\in\mathcal{B}_{s,w}} \sum_t M_{i,j,t} f_{i,j,t}}{\sum_{(i,j)\in\mathcal{B}_{s,w}} \sum_t M_{i,j,t}}. \tag{12}$$

This weights rank-local micro-batches equally, with per-token weights inversely proportional to each micro-batch's selected-token count rather than uniform globally.

Clipped sampled-token KL. For a generated token $w_{i,j,t}$, the implemented low-variance penalty is

$$\begin{aligned} u_{i,j,t} &= \log \pi_{\text{ref}}(w_{i,j,t} \mid \tau_{i,j,<t}) - \log \pi_\theta(w_{i,j,t} \mid \tau_{i,j,<t}), \\ \widehat{\text{KL}}^{\text{LV}}_{i,j,t} &= \text{clip}(\exp(u_{i,j,t}) - u_{i,j,t} - 1, -\kappa, \kappa)\,. \end{aligned} \tag{13}$$

Since $\exp(u) - u - 1 \geq 0$, the negative clipping bound is inactive; only the upper bound limits outlier token penalties and, when active, truncates their gradients.

Necessity-factorization definitions. The searched branch uses $n_i d_{i,j}$ and the direct branch uses b_i. The *w/o searched gate* variant replaces $n_i d_{i,j}$ by $d_{i,j}$; *w/o direct bonus* removes b_i; and *w/o both necessity terms* applies both changes. The shuffled control replaces only n_i by a cross-question permutation $\tilde{n}_i$, redrawn per seed, while keeping b_i paired with its original question. Thus the shuffle tests searched-branch assignment rather than the complete necessity signal or a uniform increase in reward magnitude; outcomes are reported in Table 3. On Qwen2.5-3B, the shuffled control ($0.341/0.869/3.204$) lands close to *w/o both necessity terms* ($0.339/0.861/3.286$) and below the full method ($0.368/0.972/2.812$). This ordering is consistent with value from the within-question pairing of n_i and $d_{i,j}$ rather than merely from the presence of a necessity term.

Proof and degenerate cases for Proposition 1. We restate the controlled-gate construction of Section 3.4 in full. For a fixed question i, $\gamma \in [0,1]$ replaces n_i only in the correct-searched reliance term, while b_i and all other reward branches remain fixed; $\gamma = n_i$ recovers Search-G1 and $\gamma = 1$ gives the controlled no-gate comparison, which changes only the searched reward coefficient and does not redefine b_i. Define

$$x_{i,j} = \begin{cases} d_{i,j}, & v_{i,j} = 1,\ e_{i,j} = 1,\ N_{i,j} > 0, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathbf{r}_i(\gamma) = \mathbf{a}_i + \lambda_g \gamma \mathbf{x}_i, \qquad \mathbf{a}_i := \mathbf{r}_i(0). \tag{14}$$

Let

$$H = I_G - \frac{1}{G}\mathbf{1}\mathbf{1}^\top, \qquad \mathbf{u}_i = H\mathbf{a}_i, \qquad \mathbf{v}_i = H\mathbf{x}_i,$$

and write $\mathbf{z}_i(\gamma) = \mathbf{u}_i + \lambda_g \gamma \mathbf{v}_i$. Since $\sigma_i(\gamma)$ is the sample standard deviation,

$$\sigma_i(\gamma) = \frac{\|\mathbf{z}_i(\gamma)\|_2}{\sqrt{G-1}}.$$

Proof of Proposition 1. Centering Equation (14) gives $H\mathbf{r}_i(\gamma) = \mathbf{u}_i + \lambda_g\gamma\mathbf{v}_i = \mathbf{z}_i(\gamma)$. Substituting the sample standard deviation gives Equation (10); its jth component is the scalar advantage in Equation (8). For $\delta = 0$,

$$\mathbf{A}_i^0(\gamma) = \sqrt{G-1}\,\frac{\mathbf{z}_i(\gamma)}{\|\mathbf{z}_i(\gamma)\|_2}. \tag{15}$$

Thus normalization removes centered reward scale and retains only direction. Suppose $\mathbf{u}_i$ and $\mathbf{v}_i$ are linearly independent but the centered vectors for distinct γ_1 and γ_2 have the same direction. Then for some $q > 0$,

$$\mathbf{u}_i + \lambda_g\gamma_1\mathbf{v}_i = q(\mathbf{u}_i + \lambda_g\gamma_2\mathbf{v}_i).$$

Linear independence forces $q = 1$ and then $\gamma_1 = \gamma_2$, a contradiction. If $\mathbf{u}_i$ and $\mathbf{v}_i$ are collinear, all nonzero $\mathbf{z}_i(\gamma)$ are collinear, and any global reversal between two gate values must pass through $\mathbf{z}_i(\gamma) = \mathbf{0}$. The denominator in Equation (10) is positive for $\delta > 0$, so it changes magnitude but not direction; consequently $\operatorname{sign}(A_{i,j}^\delta(\gamma)) = \operatorname{sign}(z_{i,j}(\gamma))$, and the credit of trajectory j changes sign between two gate values exactly when $z_{i,j}(\gamma_1)z_{i,j}(\gamma_2) < 0$.

For the strict homogeneous case, suppose every trajectory is valid, correct, searched, and has the same search cost c_i. Then $\mathbf{a}_i = (1-\eta c_i)\mathbf{1}$, so $\mathbf{u}_i = \mathbf{0}$ and $\mathbf{v}_i = H\mathbf{d}_i$. Let $s_{d,i}$ be the sample standard deviation of $\{d_{i,j}\}_{j=1}^G$. Equation (10) reduces componentwise to

$$A_{i,j}^\delta(\gamma) = \frac{\lambda_g\gamma(d_{i,j} - \bar{d}_i)}{\lambda_g\gamma s_{d,i} + \delta}. \tag{16}$$

For $\gamma > 0$ and $s_{d,i} > 0$,

$$\mathbf{A}_i^0(\gamma) = \frac{H\mathbf{d}_i}{s_{d,i}}, \qquad \mathbf{A}_i^\delta(\gamma) = \frac{\lambda_g\gamma s_{d,i}}{\lambda_g\gamma s_{d,i} + \delta}\mathbf{A}_i^0(\gamma). \tag{17}$$

Hence every positive gate preserves rollout ranking, advantage signs, and credit direction; it cancels from the advantage values when $\delta = 0$ and only rescales the entire vector when $\delta > 0$. □

Boundary cases. If $\gamma = 0$, the stabilized advantage in the strict homogeneous case is $\mathbf{A}_i^{\delta}(0) = \mathbf{0}$, whereas the idealized $\delta = 0$ expression is $0/0$ and undefined. If $s_{d,i} = 0$, reliance provides no within-group tie-breaking; if $\lambda_g = 0$, the gate has no effect by construction. A zero vector $\mathbf{z}_i(\gamma)$ is likewise degenerate.

Mixed groups. Outcome branches, direct-versus-search branches, and unequal search costs can make $\mathbf{u}_i \neq \mathbf{0}$, so the gate reweights $\mathbf{u}_i$ against $\mathbf{v}_i$ in Equation (10). Non-collinearity of these centered components, rather than heterogeneous composition alone, is what makes the normalized credit direction gate-dependent. For any two nonzero comparison vectors, the cosine diagnostic is exactly the cosine between $\mathbf{z}_i(\gamma_1)$ and $\mathbf{z}_i(\gamma_2)$, because the stabilizer changes only their positive scalar magnitudes.

Cosine similarity is undefined and sign-flip statistics are convention-dependent for zero comparison vectors. We therefore rely on the analytic characterization above; the controlled gate ablation in Table 3 provides complementary downstream evidence.

A.2 Experimental setup and reproduction

A.2.1 Implementation details

Data and retrieval. Unless stated otherwise, methods share policy-training data, the retriever, and the global interaction/update cap. Benchmark examples come from `RUC-NLPIR/FlashRAG_datasets` using the benchmark-native training and available held-out evaluation splits. Retrieval uses the Wikipedia 2018 corpus (`wiki-18.jsonl`), `intfloat/e5-base-v2` mean-pooled embeddings, and a FAISS Flat inner-product index, returning three passages per executed retrieval action.

Rollout and optimization. We allow four interaction rounds and maximum prompt/response/start/observation lengths of 4096/500/2048/500. Fixed-rollout methods use group size five; adaptive methods reallocate samples within the same cap. GRPO uses sample-standard-deviation normalization with $\delta = 10^{-6}$, temperature 1.0, top-$p = 0.95$, top-$k = -1$, batch size 128, learning rate 10^{-6}, and warmup ratio 0.285.

Policy update and evaluation. The clipped update uses one epoch, $\epsilon = 0.2$, mini-/micro-batches of 32/8, entropy coefficient $\alpha_{\mathrm{H}} = 0.001$, and the sampled-token KL term in Equation (13) with $\beta_{\mathrm{KL}} = 0.001$ and fixed $\kappa = 10$. Prompt, observation, and padding tokens are excluded from the actor loss. Unless a table specifies sampling or a fixed subset, single-trajectory metrics use greedy decoding on the full benchmark-specific held-out split. Each evaluation checkpoint is fixed before this final evaluation; held-out test metrics are never used to select it.

Data separation. Policy-training questions, the 300-question readout-calibration pool, validation/architecture-selection groups, and final benchmark-evaluation questions are disjoint and identified by stable item IDs. Trained methods use four independently seeded policies, whereas Base and Prompted Search are untrained point estimates.

Statistical unit. The independent training run is the unit of replication for trained-method comparisons. With four runs, a two-sided 95% t interval has half-width approximately $1.59\,\mathrm{SD}$, so small mean differences cannot support superiority by mean ordering alone. Using the seed-by-question matrices, we propagate both sources of variation. Accordingly, Table A1 resamples training seeds and, within each sampled seed, paired questions, then applies Holm correction across EM, TC, D-Inv, and Search/Q within each comparator–scale family. Search-R1 is the primary comparator, against which all four endpoints differ at both scales. The metric-wise upper-envelope rows select the strongest trained comparator by test mean separately for each endpoint and scale; they are an exploratory secondary comparison. Holm correction covers the four endpoints within each comparator–scale family, not this post-hoc comparator selection. Question resampling does not increase the four independent training replications, so the resulting small p values should not be read as precise evidence of superiority across training runs. Only TC separates at both scales, D-Inv and Search/Q separate at 7B, and the secondary analysis does not establish EM superiority over the upper envelope at either scale.

Table A1: Pairwise performance differences on NQ across training seeds and evaluation questions. Entries report hierarchical-bootstrap 95% CIs and Holm-adjusted p values. ΔEM/ΔTC are Search-G1 minus comparator; D-Inv reverses this sign; negative ΔSearch/Q favors Search-G1; †: adjusted $p < 0.05$.

Comparator	Scale	ΔEM	ΔTC	D-Inv benefit	ΔSearch/Q
Search-R1	3B	$+0.101$ $[+0.071, +0.131]$ $.004^{\dagger}$	$+0.059$ $[+0.038, +0.080]$ $.004^{\dagger}$	$+0.015$ $[+0.005, +0.025]$ $.012^{\dagger}$	-1.187 $[-1.440, -0.934]$ $.004^{\dagger}$
	7B	$+0.095$ $[+0.020, +0.170]$ $.032^{\dagger}$	$+0.049$ $[+0.030, +0.068]$ $.004^{\dagger}$	$+0.018$ $[+0.008, +0.028]$ $.008^{\dagger}$	-1.230 $[-1.460, -1.000]$ $.004^{\dagger}$
Strongest per-metric	3B	$+0.016$ $[-0.012, +0.044]$ $.320$	$+0.018$ $[+0.004, +0.032]$ $.032^{\dagger}$	$+0.004$ $[-0.003, +0.011]$ $.320$	-0.028 $[-0.114, +0.058]$ $.520$
	7B	$+0.017$ $[-0.006, +0.040]$ $.180$	$+0.018$ $[+0.006, +0.030]$ $.016^{\dagger}$	$+0.005$ $[+0.001, +0.009]$ $.048^{\dagger}$	-0.190 $[-0.286, -0.094]$ $.008^{\dagger}$

Reward and parser settings. We set $\lambda_g = 0.6$, $\eta = 0.02$, $\alpha_{\mathrm{cb}} = 0.5$, $r_{\mathrm{inv}} = -0.1$, and $r_{\mathrm{wrong}} = 0.05$, satisfying $r_{\mathrm{inv}} < r_{\mathrm{wrong}} < 1 - \eta$. The default $(\lambda_g, \eta) = (0.6, 0.02)$ is fixed from the group-split calibration/validation data, using the coefficient sweep in Table A12 and the refresh study in Table A13; test splits are not used for selection. Baseline-specific coefficients are tuned under the same validation protocol and shared budget. The search cost follows Equation (5) and is computed from the decoded policy-generated response only: $c_{i,j} = \mathbf{1}[N_{i,j} > 1] = \min\{1, (N_{i,j} - 1)_+\}$. Markers are counted on the policy-generated response only, so in-context demonstrations in the prompt are never counted. The same response-only parser is applied to direct, one-search, repeated-search, invalid-action, wrong-answer, and correct-answer trajectories.

Evaluation-time resource accounting. The evaluator records response-side marker count, policy-generated tokens, retrieved-observation tokens, and wall-clock time for every question. Policy-generated tokens are obtained from the response-side `info_mask`; observation tokens are the response-side attention-mask count minus that quantity. Table A4 reports the 3B per-question columns, and Table A5 reports the corresponding 7B audit.

Hidden-state extraction. Both readouts use the selected decoder-block output (hook on `layers[L-1]` for layer L), before the model's final normalization and without token pooling. The input for $h_{p,i}$ is the initial search-enabled prompt; the input for $h_{\mathrm{ans},i,j}$ appends the response up to, but excluding, `</answer>`, including the generated answer text. Inputs are tokenized without added special tokens, left-padded, and truncated to $4{,}096$ tokens; each state is taken at the last retained token.

Readout refresh and overhead. Readouts are selected from logistic regression and MLPs with hidden sizes 64, 128, and $(256, 128)$ on frozen states. At each refit, architecture and layer $\{21, 24, 27\}$ are selected by group-split validation AUC separately for D_ϕ and B_ψ. The nominal schedule emits checkpoints every $K = 25$ updates and refits both readouts asynchronously on $M_{\mathrm{cal}} = 300$ calibration questions. A new evaluator is installed atomically only after both readouts finish; otherwise the previous complete evaluator remains active, so the realized refresh interval may exceed the nominal cadence. Sensitivity appears in Table A13; logistic candidates use class balancing and MLPs use early stopping with $L_2 = 10^{-4}$.

Beyond the search-enabled calibration rollouts, target regeneration uses one no-retrieval decode per calibration question and one evidence-deleted decode per valid searched calibration rollout. This intervention-specific count is $M_{\mathrm{cal}} + N_{\mathrm{vs}} \le M_{\mathrm{cal}}(1 + G)$: at $M_{\mathrm{cal}} = 300$ and $G = 5$, at most 1.8k decodes, or approximately 11–12% of the $16{,}000$ rollout decodes in a nominal $K = 25$ cycle. Reward-time scoring additionally requires a batched, no-gradient teacher-forced pass through the frozen snapshot, but no LLM-judge call during policy updates.

Scoring rollouts with the hard labels instead of the readouts would change this accounting. Rewarding z_i directly needs one closed-book decode per training question, or $128 \times 25 = 3{,}200$ decodes per

cycle (20% of rollout decodes); rewarding $s_{i,j}^{\mathrm{chg}}$ directly needs one evidence-deleted decode per searched rollout, or up to $16{,}000$ decodes (100%), since every scored trajectory lies outside the calibration pool. The readouts therefore replace a per-rollout counterfactual decode with a shared forward pass, and return a graded score at the moment credit is assigned. Figure 3 and Table A17 report how closely they track the same targets.

Baseline-specific operations. The common on-policy phase uses the same policy-training data, base model, retrieval environment, trajectory format, and global interaction/update cap. Adaptive methods may reallocate rollouts within that cap, while each baseline retains its defining auxiliary computation; the comparison therefore matches shared on-policy exposure rather than end-to-end compute.

GiGPO adds anchor-state rollout groups for stepwise relative credit; AgentPRM uses a fixed external process-reward model; GPT-5 Judge scores completed trajectories with our fixed bespoke support rubric; ARPO uses post-tool entropy to trigger branch sampling and applies shared-prefix advantage attribution; IGPO rewards turn-level gains in gold-answer probability; and Search-E1 adds offline token-level forward-KL distillation from privileged sibling contexts. SAAS retains disabled/enabled rollout contrasts and boundary-aware rewards under an adapted staged schedule; β-GRPO gates correctness by minimum query-token confidence; and KbPO combines no-retrieval certainty sampling with retrieval-quality and behavior classification to form quadrant rewards. Exact auxiliary computation and shared-environment adaptations are summarized in Table A2.

Our adapted Search-R1 reference augments the original terminal-correctness reward with a fixed 0.5 correct-direct bonus and uses no necessity or reliance readout. Matching the maximum direct-answer bonus avoids comparison only against a bonus-free objective, but does not make the rewards identical: Search-G1 uses the state-dependent bonus $0.5b_i$, necessity-conditioned reliance credit, and repeated-search cost. We therefore treat Search-R1 as an outcome-plus-directness reference, not an isolated causal estimate of the readouts' effect.

Per-baseline implementation audit. Table A2 records whether a comparator is algorithm-faithful under the shared protocol (F), an adapted port retaining the published core mechanism (A), or a bespoke comparator for this study (B). These codes describe implementation provenance rather than source-code identity.

Bespoke external-judge comparator. The **GPT-5 Judge** is not a reproduction of AgentHallu. GPT-5 scores completed trajectories with a fixed evidence-support rubric and output parser designed for this study (Singh et al., 2025). AgentHallu motivates agent-trajectory hallucination assessment but does not define this reward; the judge version and prompt remain fixed across runs (Liu et al., 2026).

A.3 Results analysis and additional diagnostics

A.3.1 Evaluation definitions and diagnostics

Table A2: Baseline implementations and auxiliary computation under the shared protocol. "F" denotes an algorithm-faithful implementation under the shared protocol, "A" an adapted port retaining the published core mechanism, and "B" a bespoke comparator.

Method	Status	Auxiliary compute	Shared-protocol adaptation / provenance
Search-R1	A	none	outcome reward augmented with fixed direct bonus
GiGPO	F	anchor-state rollout groups	anchor-group size set to shared cap
AgentPRM	A	fixed process-reward-model scoring	checkpoint and prompt fixed
GPT-5 Judge	B	fixed external-judge calls	bespoke support rubric; model/prompt fixed
ARPO	F	post-tool entropy + branch rollouts	adaptive allocation within shared cap
IGPO	F	shared-model likelihood passes	confidence estimator from shared model
Search-E1	F	offline sibling-context distillation	distillation buffer within cap
SAAS	A	disabled/enabled rollout groups	boundary reward retained; schedule aligned to cap
β-GRPO	A	query-token log probabilities	confidence gate retained under the shared protocol
KbPO	A	certainty samples + retrieval/behavior judging	quadrant reward ported to shared training data

Table A3: Dataset structure and closed-book sufficiency across four QA benchmarks for Base Qwen2.5-3B-Instruct. Cells report count (percentage) on the full held-out splits. Structural labels are dataset-provided and are not interchangeable with policy-relative retrieval necessity.

Dataset	N	Exact CB-correct	Exact CB-wrong	Incl. sufficient	Incl. necessary
NQ	3,610	512 (14.18%)	3,098 (85.82%)	761 (21.08%)	2,849 (78.92%)
HotpotQA	7,405	1,134 (15.31%)	6,271 (84.69%)	1,437 (19.41%)	5,968 (80.59%)
2Wiki	12,576	2,665 (21.19%)	9,911 (78.81%)	3,115 (24.77%)	9,461 (75.23%)
MuSiQue	2,417	71 (2.94%)	2,346 (97.06%)	101 (4.18%)	2,316 (95.82%)

Dataset	Dataset structure
NQ	single-hop benchmark; no item-level hop field
HotpotQA	79.92% bridge; 20.08% comparison
2Wiki	41.63% compositional; 24.17% comparison; 12.32% inference; 21.88% bridge–comparison
MuSiQue	51.80% 2-hop; 31.44% 3-hop; 16.76% 4-hop

Table A4: Comparison of search and token costs on NQ with Qwen2.5-3B-Instruct. Evaluation uses greedy decoding on the held-out split under a matched interaction budget. Tokens are per-question means rounded to integers; observation tokens are bounded by the per-round observation cap, so their spread across methods is compressed relative to marker counts. Table A5 reports the corresponding 7B audit.

Method	EM↑	Search/Q	Gen. tok/Q	Obs. tok/Q
Base	0.142	0.000	30	0
Prompted Search	0.250	2.215	120	500
Search-R1	0.352	3.281	149	510
Search-G1	0.453	2.094	**85**	**468**

Table A5: Comparison of search and token costs on NQ with Qwen2.5-7B-Instruct. Evaluation uses greedy decoding on the held-out split under a matched interaction budget, as in the 3B audit in Table A4. Tokens are per question.

Method	EM↑	Search/Q	Gen. tok/Q↓	Obs. tok/Q↓
Base	0.212	0.000	28	0
Prompted Search	0.352	2.171	110	490
Search-R1	0.417	3.090	140	500
Search-G1	0.512	1.860	**78**	**440**

Grounding metric definitions. TC's searched-trajectory component and D-Inv use a document-level counterfactual intervention on originally correct searched trajectories. The evaluator replaces the gold answer in answer-bearing retrieved documents with an alternative entity or value, then regenerates an answer from the modified documents. Every evaluated correct searched trajectory is assigned exactly one class: CF-FOLLOW when its normalized answer changes to track the injected counterfactual entity or value, INVARIANT when it remains invariant and preserves the original correct answer, and OTHER otherwise. This evaluation protocol is distinct from the evidence-deletion labels used to fit the evidence-reliance readout: the training-time label perturbs evidence by *deletion* and is used only to fit $D_{\phi^{(m)}}$, while TC/D-Inv perturb evidence by *document-level entity replacement* and are computed post hoc without feeding back into the reward.

Proxy scope and triangulation. The deletion intervention retains the reasoning prefix and compares a sampled realized answer with a deterministic counterfactual decode, thereby measuring answer-stage sensitivity conditional on that prefix. Table A6 reports the stricter prefix-deletion audit, while Table A15 reports paired-decoding controls. Document replacement supplies a non-identical held-out intervention with no shared gradient path, while the independent human evaluation uses neither perturbation nor learned readout. Agreement across these views tests whether the grounding pattern persists across distinct operationalizations of evidence use rather than being inherited from the training label.

For a search-enabled policy with at least one classified correct searched trajectory, let $\mathcal{C}$ be the union of originally correct direct trajectories and originally correct searched trajectories classified by the intervention. **TC** is a graded score of correct-direct retention and counterfactual response

$$\mathrm{TC} = \frac{1}{|\mathcal{C}|} \sum_{\tau \in \mathcal{C}} s(\tau), \qquad s(\tau) \in \{1.0, 0.6, 0.3\}.$$

Correct direct trajectories receive score 1.0 as the selective-answering component; classified correct searched trajectories receive 1.0 for CF-FOLLOW, 0.6 for OTHER, and 0.3 for INVARIANT. Thus TC is a joint search-policy measure of correct direct retention and counterfactual response. Because its absolute level depends on fixed heuristic weights and the direct-answer convention, we interpret it jointly with D-Inv, a weight-free searched-only proportion, and with the S-TC control reported in Table A14, which removes direct trajectories from the denominator. Search-G1 retains the leading value under these companion endpoints, so its qualitative advantage is not driven solely by TC's direct-answer convention. TC and D-Inv are undefined for the search-disabled Base reference because no retrieved context is available for the intervention protocol. **D-Inv** is counterfactual-document invariance among classified originally correct searched trajectories,

$$\text{D-Inv} = \frac{N_{\text{Invariant}}}{N_{\text{CF-follow}} + N_{\text{Invariant}} + N_{\text{Other}}},$$

so lower is better. D-Inv contains no direct-answer term and therefore avoids TC's mechanical direct-answer contribution, although it remains conditional on each policy's correct searched trajectories. Table 1 shows the NQ comparison, and Table 2 shows the same D-Inv/TC ordering on 2Wiki. We read D-Inv as the primary evidence-dependence endpoint and TC as a summary of correct-direct retention and counterfactual response; S-TC provides the complementary searched-only control for TC's composition effect.

Document replacement is an operational diagnostic of counterfactual evidence dependence, not a necessary-and-sufficient test of grounding. Under answer-bearing-fact replacement, invariance is evidence of weak dependence on the manipulated fact, whereas responsiveness may reflect evidential reasoning, surface context following, or perturbation-induced failure. Moving an invariant answer to OTHER also lowers D-Inv. We therefore interpret D-Inv jointly with the measured surface controls in Table A16, the independent human support evaluation in Table A10, and the common-question analysis in Section A.3.4. Unless noted, Table 1 reports TC and D-Inv on NQ.

Readout-audit provenance. The NQ–3B representation result in Figure 3 uses layer 27 on $3{,}663$ correct trajectories ($2{,}015$ sensitive; $1{,}648$ invariant), separate from the scalar/logit-control pool. Figure 5 uses Qwen2.5-3B/7B-Instruct on 2WikiMultiHopQA; the 3B series covers checkpoints $0, 25, 50, 75, 100, 150$.

Table A6: Agreement between default and strict evidence-deletion labels on NQ and HotpotQA at 3B and 7B.

Dataset	Agreement↑	κ↑	Reliance Rate	Strict Reliance Rate
Qwen2.5-3B-Instruct				
NQ	0.953	0.881	0.724	0.747
HotpotQA	0.952	0.879	0.712	0.741
Qwen2.5-7B-Instruct				
NQ	0.956	0.897	0.681	0.704
HotpotQA	0.954	0.894	0.668	0.696

Layer-wise linear separability. We further examine where the deletion-sensitivity label is most linearly separable. For each layer, we train the same linear readout to distinguish deletion-sensitive from deletion-invariant correct trajectories and compare separability across depth. As shown in Figure A1, the best layer varies across settings—layer 27/21 on NQ and 24/21 on 2Wiki for 3B/7B—supporting layer selection by model and refit rather than one universal depth.

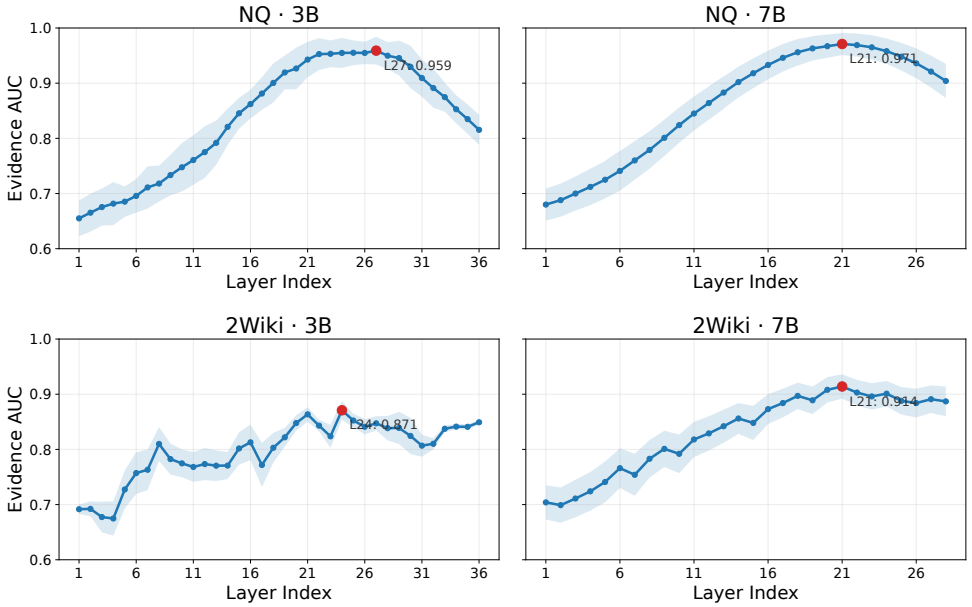


Figure A1: Layer-wise linear separability of evidence-deletion sensitivity on NQ and 2WikiMultiHopQA using Qwen2.5-3B and Qwen2.5-7B. Each point reports commit-token readout AUC among $\mathrm{EM} = 1$ rollouts.

Counterfactual label reliability. As defined in Equation (1), Search-G1 operationalizes evidence reliance by whether the realized answer changes under evidence ablation. The 3B and 7B NQ/HotpotQA rows compare the default deletion label against a stricter intervention that also removes intermediate reasoning text. Agreement is approximately 0.95 with $\kappa \approx 0.88$, showing that reasoning-prefix leakage is limited in the audited settings.

Metric definitions for the closed-book analyses. We define the HotpotQA metrics in Table A7 and the 2Wiki metrics in Table A8. The former defines CB-correct separately per policy from its deterministic retrieval-disabled decode; the latter fixes Base-model CB-wrong within each scale. **CB-EM/CBw-EM** are exact match on these conditional sets, while **CB-S/Q/All-S/Q** count response-side search markers per question on the CB-correct/full audited slice. **Over-S** is the valid-trajectory fraction with multiple markers; **Any@3** is the evaluated CB-wrong fraction with at least one correct sampled trajectory. These metrics condition on valid parsed answers, so counts may differ across methods. **Valid-S/Q** names this conditional 2Wiki statistic; Tables 2 and A14 use all-question Search/Q. None uses evidence perturbation.

Table A7: Known-answer preservation and search behavior on a 1,000-question HotpotQA slice. CB-correct is defined separately for each policy from its retrieval-disabled decode.

Method	CB-EM↑	CB-S/Q↓	All-S/Q↓	Over-S↓
Qwen2.5-3B-Instruct				
Prompted Search	0.380	1.199	1.215	0.225
Search-R1	0.541	1.535	1.629	0.489
Search-G1	**0.573**	**1.056**	**1.049**	**0.040**
Qwen2.5-7B-Instruct				
Prompted Search	0.431	1.085	1.102	0.194
Search-R1	0.592	1.411	1.503	0.438
Search-G1	**0.641**	**0.914**	**0.905**	**0.028**

Table A8: Retrieval performance on fixed Base-model closed-book-wrong subsets of 2WikiMultiHopQA. Metrics condition on valid trajectories; methods share question IDs within each scale.

Method	CBw-EM↑	Any@3↑	Valid-S/Q↓	Over-S↓
Qwen2.5-3B-Instruct				
Prompted Search	0.145	0.236	1.129	0.315
Search-R1	0.254	0.333	1.972	0.809
Search-G1	**0.327**	**0.435**	1.649	0.605
Qwen2.5-7B-Instruct				
Prompted Search	0.162	0.252	1.050	0.284
Search-R1	0.278	0.362	1.840	0.748
Search-G1	**0.356**	**0.476**	1.520	0.548

Search-action frequency and known-answer preservation. We test whether retrieval training preserves answers within each policy's estimated knowledge boundary. Table A7 shows that, on the audited HotpotQA slice, Search-G1 has the highest CB-EM and the lowest CB-S/Q and All-S/Q among the compared methods. Relative to Search-R1, CB-EM is 0.573 versus 0.541 at 3B and 0.641 versus 0.592 at 7B; All-S/Q is 1.049 versus 1.629 and 0.905 versus 1.503. Because each method defines its own CB-correct set, the following analysis complements this view with a fixed Base-model CB-wrong set on 2Wiki.

Closed-book-wrong question performance. We next evaluate the CB-wrong subset, on which the base model fails without retrieval. Table A8 shows that, on 2WikiMultiHopQA, Search-G1 reaches mean valid-conditioned EM $0.327/0.356$ versus $0.254/0.278$ for Search-R1 at 3B/7B, with Valid-S/Q $1.649/1.520$ versus $1.972/1.840$. Thus, in this subset, lower mean search-marker frequency co-occurs with higher mean valid-conditioned EM.

Cross-dataset transfer. Table A9 maps each source-trained Search-G1 policy to all four held-out targets. Across scales, $23/24$ off-diagonal means exceed Prompted Search. For NQ-trained Search-G1 versus NQ-trained Search-R1, paired EM differences [95% CI] are $+0.101\,[+0.072, +0.130]$, $+0.008\,[-0.012, +0.028]$, $+0.006\,[-0.015, +0.027]$, and $+0.001\,[-0.012, +0.014]$ on NQ, HotpotQA, 2Wiki, and MuSiQue, respectively. The largest and clearest gain is in-domain; the smaller out-of-domain differences have intervals spanning zero.

Independent human grounding evaluation. Section 4.3 summarizes this evaluation; here we give the per-method table, the annotation protocol, and the agreement statistics. Each pool contains 180 correct searched instances drawn independently per method, so answer correctness is fixed by construction and annotators judge only whether the retrieved evidence supports the answer.

Annotation. Three annotators independently label each instance on a 3-point support scale: (i) **Fully supported** (2): the evidence explicitly states, or unambiguously entails through a short and complete reasoning chain, every fact required for the answer; (ii) **Partially supported** (1): the evidence

Table A9: In-domain and cross-dataset transfer of Search-G1 across four QA benchmarks at 3B and 7B. Each cell reports EM / Search/Q; diagonal cells are in-domain and off-diagonal cells are cross-dataset. Grounding comparisons focus on NQ and 2Wiki in Tables 1 and 2.

Training Dataset	Target Test Dataset			
	NQ	HotpotQA	2Wiki	MuSiQue
Qwen2.5-3B-Instruct				
Prompted Search	0.250 / 2.215	0.199 / 2.531	0.238 / 2.951	0.057 / 2.141
NQ	**0.453 / 2.094**	0.248 / 2.480	0.256 / 2.551	0.091 / 2.406
HotpotQA	0.359 / 2.205	**0.306 / 2.549**	0.270 / 2.828	0.099 / 2.881
2Wiki	0.377 / 2.486	0.316 / 2.439	**0.368 / 2.812**	0.116 / 3.275
MuSiQue	0.297 / 3.640	0.241 / 3.389	0.279 / 3.227	**0.143 / 3.443**
Qwen2.5-7B-Instruct				
Prompted Search	0.352 / 2.171	0.292 / 2.480	0.274 / 2.892	0.074 / 2.098
NQ	**0.512 / 1.860**	0.365 / 2.202	0.330 / 2.265	0.162 / 2.137
HotpotQA	0.406 / 1.958	**0.450 / 2.263**	0.349 / 2.511	0.176 / 2.558
2Wiki	0.426 / 2.208	0.465 / 2.166	**0.475 / 2.495**	0.206 / 2.908
MuSiQue	0.336 / 3.232	0.354 / 3.009	0.360 / 2.866	**0.254 / 3.057**

Table A10: Human evaluation of answer support on independent 180-instance correct-searched pools per method with Qwen2.5-3B. S@C is mean normalized support; Full and Uns./CBU are the fully supported and unsupported fractions. This evaluation uses no evidence perturbation, re-decoding, or learned readout.

Method	S@C (↑)	Full (↑)	Uns./CBU (↓)
Prompted Search	0.594	0.400	0.211
Search-R1	0.650	0.472	0.172
GPT-5 Judge	0.797	0.678	0.083
Search-E1	0.719	0.561	0.122
w/o Evidence-Reliance Reward	0.633	0.450	0.183
Search-G1 (ours)	**0.842**	**0.733**	**0.050**

supports only part of the answer, covers only some hops, or is relevant but insufficient to entail the answer; (iii) **Unsupported** (0): the evidence is irrelevant, contradicts the answer, or does not contain enough information to support the answer. The final label is the majority vote; all-disagreement cases are adjudicated by a fourth senior annotator.

Target measures. We report **Support@Correct** (S@C), the mean majority-vote support score normalized to $[0, 1]$; **Fully Supported** (Full), the fraction of score-2 instances; and **Unsupported** (Uns.), the fraction of score-0 instances. Since all evaluated trajectories are exact-match correct, Uns. is also the *Correct-but-Unsupported* (CBU) rate: the fraction of correct answers not supported by the retrieved evidence shown to annotators. This evaluation uses no evidence-deletion labels, counterfactual document replacement, re-decoding, or Search-G1 reward readouts.

Agreement. The annotation reaches substantial pre-adjudication agreement (Fleiss' $\kappa = 0.79$, ordinal Krippendorff's $\alpha = 0.81$ on Qwen2.5-3B-Instruct).

Table A10 compares six methods on distinct correct-searched pools. The unconditional common-question decomposition in Section A.3.4 provides a complementary view on common question IDs.

A.3.2 Sensitivity and controlled comparisons

Shared anchors and variant definitions. The component, coefficient, refresh, and unified-accounting studies below reuse the same four trained policies as comparison anchors rather than reporting additional repetitions. The $\eta = 0$ row is the *w/o search cost* condition, $K = \infty$ is the no-refresh evaluator condition, and *w/o searched gate* is the main no-gate condition. In the latter, n_i is

Table A11: Hyperparameter candidates and validation-selected settings for Search-G1 and baseline methods. Candidate counts and per-method selections use the shared validation rule; Search-G1's default matches the reported coefficient/update sweeps. "Scale" is relative to the official auxiliary-reward coefficient.

Method	Tuned quantities	Configs per scale	Selected setting (3B / 7B)
Search-G1	$\lambda_g, \eta, K, M_{\text{cal}}$	12	$(0.6, 0.02, 25, 300)$ / same
Search-R1	direct bonus; KL coefficient	9	$(0.5, 10^{-3})$ / same
GiGPO	state-credit scale; anchor-group cap	9	$(1.0, 5)$ / same
AgentPRM	process-reward scale; KL coefficient	9	$(1.0, 10^{-3})$ / same
GPT-5 Judge	support-reward scale; KL coefficient	9	$(0.5, 10^{-3})$ / same
ARPO	entropy scale; allocation temperature	9	$(1.0, 1.0)$ / same
IGPO	information-gain scale; normalization	9	$(1.0, \text{group})$ / same
Search-E1	distillation mixture; temperature	9	$(0.25, 1.0)$ / same
SAAS	boundary threshold; allocation scale	9	$(0.5, 1.0)$ / same
β-GRPO	confidence threshold; reward scale	9	$(0.5, 1.0)$ / same
KbPO	certainty weight; quality weight	9	$(0.5, 0.5)$ / same

Table A12: Sensitivity to evidence-reliance and search-cost coefficients on 2WikiMultiHopQA.

λ_g	η	EM↑	TC	Search/Q↓
Qwen2.5-3B-Instruct				
0.4	0.02	0.347	0.854	2.842
0.6	0.02	0.368	0.972	2.812
0.8	0.02	0.381	0.911	3.383
0.6	0.00	0.338	0.920	3.576
0.6	0.05	0.334	0.891	3.568
Qwen2.5-7B-Instruct				
0.4	0.02	0.457	0.878	2.521
0.6	0.02	0.475	0.978	2.495
0.8	0.02	0.486	0.926	2.980
0.6	0.00	0.449	0.934	3.144
0.6	0.05	0.446	0.909	3.138

replaced by 1 only in the correct-searched reliance term; b_i and all other branches remain unchanged. The four-round Search-G1 entry is the default interaction-budget anchor.

Validation-based hyperparameter selection. Table A11 summarizes Search-G1 coefficient and refresh sweeps together with the baseline-by-baseline tuning ledger. The same validation rule is used for every method: retain configurations within 0.02 EM of the best validation EM, maximize TC, then minimize Search/Q; test data are never consulted.

Reward coefficient sensitivity. We study sensitivity to the evidence-reliance and search-cost coefficients on 2WikiMultiHopQA. Table A12 varies λ_g and η one at a time while keeping the other training settings fixed. Among tested values, the default $(0.6, 0.02)$ gives the highest TC and lowest Search/Q at both scales; $\lambda_g = 0.8$ gives higher EM but worse TC and Search/Q, and increasing η to 0.05 increases Search/Q in these runs. We fix $\alpha_{\text{cb}} = 0.5$, matching the maximum correct-direct bonus of the Search-R1 reference. Equation (7) characterizes the local pre-normalization incentive; after retraining, policy-level Search/Q need not vary monotonically with an individual coefficient.

Estimator update sensitivity. We study how often the Search-G1 estimators should be refit and how many refitting questions are needed at each update. As the policy changes, the targets and trajectory-conditioned state distributions move away from the initial calibration data. Table A13 evaluates different co-evolution schedules at both model scales on 2WikiMultiHopQA, where K is the refitting interval and M_{cal} is the number of refitting questions. Among tested schedules, $K = 25, M_{\text{cal}} = 300$ gives the highest EM and TC; $K = 50$ lowers Search/Q with slightly lower outcomes, while $K = 10$ or $M_{\text{cal}} = 100$ gives a weaker joint profile.

Table A13: Sensitivity to readout refresh frequency and calibration-set size on 2WikiMultiHopQA.

Update Schedule	K	M_{cal}	EM↑	TC	Search/Q↓
Qwen2.5-3B-Instruct					
No-refresh evaluator	∞	0	0.357	0.906	3.225
Nominal 50-update cadence	50	300	0.360	0.937	2.656
Nominal 25-update cadence	25	300	0.368	0.972	2.812
Nominal 10-update cadence	10	300	0.333	0.895	3.589
Nominal 25-update cadence	25	100	0.339	0.912	3.839
Qwen2.5-7B-Instruct					
No-refresh evaluator	∞	0	0.466	0.922	2.846
Nominal 50-update cadence	50	300	0.468	0.948	2.362
Nominal 25-update cadence	25	300	0.475	0.978	2.495
Nominal 10-update cadence	10	300	0.445	0.913	3.155
Nominal 25-update cadence	25	100	0.450	0.927	3.368

Common-subset comparison under unified accounting. We score all methods on one fixed 1,024-question 2WikiMultiHopQA subset under unified accounting: identical question IDs, invalid trajectories charged EM=0, and Search/Q and per-marker EM computed over all questions. To avoid overloading the main TC definition, we denote the searched-trajectory-only counterfactual-response score by **S-TC**; it averages the same counterfactual scores over correct searched trajectories only. Search-G1 attains the highest all-question EM, S-TC, and EM per search marker. This table aligns question IDs and marker accounting. Table A4 gives the matched 3B NQ token audit, and Table A5 gives the 7B counterpart.

Table A14: Comparison under unified accounting on a fixed 1,024-question 2WikiMultiHopQA subset with Qwen2.5-3B. S-TC is the searched-trajectory-only counterfactual-response score and is distinct from the union-of-correct TC. Invalid trajectories are charged EM= 0; Search/Q is averaged over all 1,024 IDs.

Method	EM↑	S-TC↑	Search/Q↓	EM per marker↑
Prompted Search	0.231	0.608	1.602	0.144
Search-R1	0.285	0.534	1.942	0.147
Search-G1	**0.355**	**0.655**	**1.580**	**0.225**

A.3.3 CASE STUDY

We provide selected qualitative question-level summaries from Qwen2.5-3B-Instruct on 2WikiMultiHopQA. Each case summarizes representative outputs for a question on which both Prompted Search and the Search-R1 baseline fail while Search-G1 answers correctly. The cases illustrate method-level differences in task accuracy, evidence alignment, and search behavior under this selection rule.



Figure A2: Qualitative comparison of selected search trajectories and evidence use on 2WikiMultiHopQA with Qwen2.5-3B-Instruct. The Evidence column gives a reference chain for interpreting each question; method rows summarize representative observed outputs rather than implying identical retrieved contexts.

A.3.4 ADDITIONAL VALIDITY AND ROBUSTNESS CHECKS

Evidence-reliance label audit. The evidence-reliance label in Equation (1) compares a sampled trajectory answer with a deterministic evidence-ablated decode, so it can mix evidence sensitivity with decoding mismatch. We therefore re-derive the label under (i) *paired-greedy* decoding and (ii) *common-random-number (CRN) paired sampling*, which couples the sampling variates to reduce independent decoding noise. On Qwen2.5-3B-Instruct/NQ at checkpoint 50, the original labels agree with these alternatives on 93%/96% of trajectories ($\kappa = 0.85/0.90$). As reported in Table A15, the positive-label rate changes by $-0.02/-0.01$, while 6%/4% of group-relative advantage signs flip. Thus decoding mismatch is not the dominant source of the label in this audit, although its optimization effect is nonzero.

Table A15: Agreement of evidence-reliance labels under paired decoding controls on NQ with Qwen2.5-3B-Instruct. Controls use paired-greedy decoding and common-random-number paired sampling.

Construction	Agree.↑	κ ↑	Pos.-rate Δ	Adv.-sign flip↓
Paired-greedy	0.93	0.85	-0.02	0.06
CRN paired sampling	0.96	0.90	-0.01	0.04

Shortcut-control audit. Table A16 reports supplementary readout diagnostics using surface features, shuffled labels, and matching on search count and response length.

Table A16: Evidence-reliance prediction under surface-feature and label-shuffling controls on NQ and 2WikiMultiHopQA with Qwen2.5-3B-Instruct.

Predictor/control	NQ AUC	2Wiki AUC
Search count only	0.548	0.562
Response length only	0.557	0.574
Retrieved-token count only	0.536	0.551
Answer–evidence lexical overlap	0.589	0.603
Label-shuffled representation	0.501	0.498
Full answer-commit representation	0.959	0.871
Matched on search count and length	0.934	0.842

Retrieval-necessity label audit. The necessity target z_i in Equation (2) comes from a deterministic closed-book decode scored by normalized answer inclusion, which is a more permissive criterion than the task exact-match verifier. On the same full held-out Base-model decodes used in Table A3, exact-match correctness is a strict subset of inclusion sufficiency, so the two labels agree on 93.10% of NQ and 96.42% of 2Wiki examples and disagree on $249/3{,}610$ and $450/12{,}576$ respectively.

Relative to EM labeling, inclusion assigns closed-book sufficiency to an additional 6.90% of NQ questions. These disagreements shift calibration targets toward sufficiency. When reflected in higher fitted b_i, the broader sufficiency labels favor direct-answer credit under Equation (7). Table A7 characterizes policy-relative CB-correct slices, and Table A8 provides the fixed Base-CB-wrong view. We use inclusion because it is robust to free-form closed-book decodes and report its agreement and disagreement counts with EM for transparency.

Necessity-score distribution across refits. We recompute $n_i = 1 - b_i$ by applying each refitted boundary readout to its matched hidden-state dump on the same 300-question 2Wiki calibration slice. Table A17 shows a strongly bimodal distribution across refits: the 10th percentile stays near zero, the median above 0.93, and the 90th percentile near one. The mean stays within 0.605–0.622 and the predicted-necessary fraction within 59.67–61.67%. The aggregate boundary composition is therefore stable across these refits, while each checkpoint uses a newly calibrated readout.

Common-question unconditional behavior decomposition. TC and D-Inv condition on method-specific correct trajectories, so an unconditional common-question analysis is needed to expose selection effects. Table A18 evaluates all five methods on the same fixed set of 600 NQ question IDs, using the full/partial/unsupported support categories defined in the preceding annota-

Table A17: Retrieval-necessity score distributions across refits on a fixed 300-question 2WikiMultiHopQA calibration slice with Qwen2.5-3B-Instruct. Hard S/N counts closed-book-sufficient/necessary labels; Pred. N is the number of readout scores with $n_i \geq 0.5$.

Checkpoint	Hard S/N	Mean	SD	q10 / median / q90	Pred. N (%)	AUC
25	115/185	0.617	0.454	0.011 / 0.969 / 1.000	185 (61.67%)	0.866
50	118/182	0.622	0.430	0.026 / 0.932 / 1.000	183 (61.00%)	0.888
75	122/178	0.605	0.448	0.013 / 0.948 / 1.000	179 (59.67%)	0.881

tion protocol. It keeps partial support separate and reports unconditional grounded accuracy $\mathrm{UGA} = P(\text{correct direct}) + P(\text{correct search, full}) + 0.5P(\text{correct search, partial})$.

Table A18: Unconditional decomposition of answer correctness and support on 600 shared NQ questions. All five methods use the same sample. Search-full, search-partial, and search-unsupported use the support categories defined in the preceding annotation protocol. The five buckets sum to one; UGA is reported with a question-bootstrap 95% CI.

Method	Corr.-direct	Search-full	Search-partial	Search-unsupp.↓	Incorrect	UGA [95% CI]↑
Prompted Search	0.10	0.09	0.05	0.05	0.71	$0.215\,[0.184, 0.248]$
Search-R1	0.09	0.15	0.07	0.06	0.63	$0.275\,[0.240, 0.312]$
GPT-5 Judge	0.11	0.21	0.07	0.03	0.58	$0.355\,[0.317, 0.394]$
Search-E1	0.11	0.19	0.07	0.04	0.59	$0.335\,[0.298, 0.373]$
Search-G1	0.12	0.20	0.08	0.03	0.57	$0.360\,[0.322, 0.399]$

On this shared question set, Search-G1 reaches unconditional grounded accuracy 0.360, compared with 0.355 for GPT-5 Judge, with the same rounded unsupported rate. This provides an independent common-denominator check that does not use Search-G1's learned reward readouts.

Credit–support association audit. Within each original $G = 5$ group, we permute $d_{i,j}$ only among correct searched trajectories with identical $c_{i,j}$, keeping other reward terms fixed. Cost-only removes only the correct-searched reliance bonus. Pearson correlations with human support use correct searched trajectories and advantages from complete original groups, with within-group centering. Table A19 reports these auxiliary associations and task outcomes.

Table A19: Association between learning credit and human-assessed support on 2WikiMultiHopQA with Qwen2.5-3B-Instruct. Support and correlations include 95% CIs.

Reward variant	Eval. N	Support N	EM↑	Human support↑	Adv.–support corr.↑
Score-permutation control	600	180	0.350	$0.630\,[0.580, 0.680]$	$0.02\,[-0.13, 0.17]$
Cost-only	600	180	0.350	$0.660\,[0.610, 0.710]$	$0.08\,[-0.07, 0.23]$
Search-G1	600	180	0.370	$0.730\,[0.680, 0.780]$	$0.41\,[0.28, 0.52]$