HindSearch: Trajectory-Level Hindsight Critique for Search-Augmented Reinforcement Learning

Anonymous ACL submission

Abstract

Search-augmented LM agents are typically trained with a binary exact-match reward, which signals only whether a trajectory succeeded without telling us why it failed. We introduce HindSearch, a hindsight self-distillation procedure for GRPO: after each rollout, a frozen judge writes a short critique of every failed trajectory using the gold answer, and the critique supplies an auxiliary on-policy distillation signal on the student's search actions. On the standard seven-benchmark suite with Qwen2.5-3B-Instruct, HindSearch reaches **39.4**% average EM, outperforming prior search-RL baselines. Removing the judge's access to the gold answer erases most of the gain, isolating hindsight as the source of the improvement. Code is available at https://anonymous.4open.science/r/hindsearch-anon-EBDC.

1 Introduction

Search-augmented LM agents interleave retrieval with reasoning and are commonly trained end-to-end under a scalar outcome reward (Jin et al., 2025; Zhao et al., 2025; Shi et al., 2025; Gao et al., 2025; Sun et al., 2026). The setup is simple but throws away signal. A failed trajectory contains the queries the agent issued, the snippets they returned, and—at training time—the gold answer; together these often pin down what should have been searched for instead. An outcome-only objective ignores all of this and asks the agent to recover the underlying search policy from a bandit signal, which matches the long credit-assignment delay and the stability issues observed in publicly released search-RL training runs.

Prior work on denser signals divides roughly into process reward models for retrieval (Wang et al., 2026b; Xia et al., 2026; Wen et al., 2026; Shu et al., 2026), which assign scalar quality scores to individual steps, and hint-conditioned distillation methods such as OpenClaw-RL (Wang et al., 2026a), which translates a process reward model's natural-language hint into token-level supervision. Process rewards remain evaluative—they grade a step rather than say what to do instead—and OpenClaw-RL's hints are produced online from the next state, without the gold answer. It is also validated on GUI control and tool-call domains, where each step has a clear local notion of success. None of this exploits the cleanest signal we have at training time for search: the answer itself.

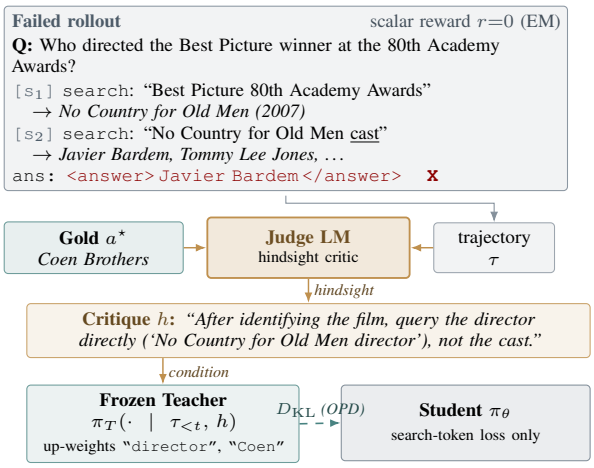

Figure 1: **HindSearch at a glance.** Standard GRPO collapses a failed multi-turn rollout (top) into a single r=0 scalar, discarding *how* the agent erred. HindSearch feeds the full trajectory τ and the gold answer $a^\star$ to a frozen judge, which emits a directive critique h. The critique conditions a frozen self-teacher whose action distribution is distilled into the student via on-policy KL (OPD), masked to search-action tokens, recovering rich corrective signal from rollouts that would otherwise produce no reward gradient.

We propose **trajectory-level hindsight critique (TLHC)**: after each GRPO rollout, every failed trajectory is sent to a frozen judge together with its gold answer, and the judge returns a short critique of what the agent should have done differently. The critique conditions a frozen teacher whose log-probabilities supply an auxiliary on-policy distil-

lation (OPD) (Agarwal et al., 2024) loss on the student's search-action tokens. Because the judge sees the answer, the critique is directive rather than evaluative; because it is produced post hoc, no online process reward model is needed. Section 3 gives the full procedure.

Our contributions are:

- **TLHC:** a post-hoc hindsight-critique procedure that converts trajectory-level EM failures into per-token OPD supervision on top of GRPO.
- **Empirical: 39.4%** average EM on the standard seven-benchmark search-QA suite, outperforming prior search-RL baselines, with stable training throughout the 300-step budget.

2 Related Work

RL for search agents and process rewards. Search-R1 (Jin et al., 2025) introduced the multi-turn search-RL recipe — PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024) under a binary EM reward — that subsequent work adopts as a common starting point. Follow-up work explores variants of this recipe: outcome-only setups (Song et al., 2025; Chen et al., 2025), multi-reward shaping (Zhao et al., 2025), refinement steps (Shi et al., 2025), longer horizons (Gao et al., 2025; Xi et al., 2025), decoupling searcher from generator (Jiang et al., 2025), knowledge-boundary-aware gating (Huang et al., 2025), simulated retrieval for cheaper training (Sun et al., 2026), live-web environments (Zheng et al., 2025), and parallel multimodal search (Li et al., 2026). A separate line introduces step-level process rewards: information-gain shaping (Wang et al., 2025; Liang et al., 2026a), turn-level credit assignment (Ma et al., 2026a), and other learned step scorers (Wang et al., 2026b; Xia et al., 2026; Wen et al., 2026; Shu et al., 2026; Cui et al., 2025). These signals are denser than EM but still evaluative—they grade a step rather than instruct a correction.

Hindsight and hint-conditioned distillation. Hindsight experience replay (Andrychowicz et al., 2018) relabels trajectories with achieved goals; AgentHER (Ding, 2026) and ECHO (Hu et al., 2026) adapt the idea to LM agents via off-policy rewriting. Concurrent work on self-distillation for search agents includes SD-Search (Ma et al., 2026b) and Search-E1 (Liang et al., 2026b). The closest prior method is OpenClaw-RL (Wang et al., 2026a), which conditions a frozen same-family teacher on a natural-language hint and distills the resulting log-probability gap into the student. We adopt its hint-prefix teacher mechanism but differ in two ways. Beyond the post-hoc-vs-online hint source already noted in §1, OpenClaw-RL folds the directive signal into a PPO-clipped surrogate restricted to a top-$K = 4$ vocabulary subset, whereas we use a separate, plain-clamp OPD loss on the full response distribution, gated to failed trajectories and masked to search-action tokens. We ablate the loss form in §5.

3 Method

3.1 Problem setup

We study multi-turn search-augmented QA with a single LM policy π_θ. At each turn t, the policy emits either a search action $\langle\text{SEARCH}\rangle q_t \langle/\text{SEARCH}\rangle$, whose retrieved result o_t is appended to the context as an $\langle\text{INFORMATION}\rangle$ block, or a final answer $\langle\text{ANSWER}\rangle a \langle/\text{ANSWER}\rangle$. A trajectory $\tau = (q_1, o_1, \ldots, a)$ receives a binary reward $r(\tau) = \mathbb{I}[\text{EM}(a, a^\star)]$ against gold answer $a^\star$.

3.2 GRPO with search rollouts

We optimize π_θ with GRPO (Shao et al., 2024) under the standard search-RL recipe (Jin et al., 2025). For each prompt, G trajectories are sampled, group-normalized advantages are computed from outcome rewards, and retrieved $\langle\text{INFORMATION}\rangle$ tokens are masked from the policy gradient. Let $L_{\text{PPO}}(\theta)$ denote the standard clipped policy-gradient objective. HindSearch adds a single auxiliary term on top.

3.3 Trajectory-level hindsight critique

After each rollout batch, we partition trajectories into $\mathcal{T}_+ = \{\tau : r(\tau) = 1\}$ and $\mathcal{T}_- = \{\tau : r(\tau) = 0\}$. For each $\tau \in \mathcal{T}_-$, a frozen judge π_J is queried with the rendered trajectory and the gold answer to produce a 1–2 sentence critique:

$$h(\tau) \;=\; \pi_J\big(\tau,\, a^\star\big). \tag{1}$$

The teacher π_{ref} is the student's frozen initialization (also reused as the GRPO KL anchor). For every search step t in τ, the hint is broadcast as a prefix:

$$c_t^T \;=\; \langle\text{HINT}\rangle\, h(\tau)\, \langle/\text{HINT}\rangle \;\oplus\; c_t. \tag{2}$$

Method	NQ†	TriviaQA	PopQA	HotpotQA†	2Wiki	MuSiQue	Bamboogle	Avg
Direct Inference	10.6	28.8	10.8	14.9	24.4	2.0	2.4	13.4
CoT (Wei et al., 2022)	2.3	3.2	0.5	2.1	2.1	0.2	0.0	1.5
IRCoT (Trivedi et al., 2023)	11.1	31.2	20.0	16.4	17.1	6.7	24.0	18.1
Search-o1 (Li et al., 2025)	23.8	47.2	26.2	22.1	21.8	5.4	**32.0**	25.5
RAG (Lewis et al., 2021)	34.8	54.4	38.7	25.5	22.6	4.7	8.0	27.0
SFT	24.9	29.2	10.4	18.6	24.8	4.4	11.2	17.6
R1-instruct (Guo et al., 2025)	21.0	44.9	17.1	20.8	27.5	6.0	19.2	22.4
Rejection Sampling	29.4	48.8	33.2	24.0	23.3	5.9	21.0	26.5
Search-R1-Instruct (Jin et al., 2025)	39.7	56.5	39.1	33.1	31.0	12.4	23.2	33.6
Search-R1-Base (Jin et al., 2025)	42.1	58.3	41.3	29.7	27.4	6.6	12.8	31.2
ReSearch-Instruct (Chen et al., 2025)	36.5	57.1	39.5	35.1	27.2	9.5	26.6	33.1
ReSearch-Base (Chen et al., 2025)	42.7	59.7	43.0	30.5	27.2	7.4	12.8	31.9
HindSearch	**45.0**	**60.5**	**45.2**	**39.5**	**40.1**	**14.2**	31.2	**39.4**

Table 1: Performance comparison across seven QA benchmarks (Qwen2.5-3B backbone, Instruct unless suffixed -Base). Best results in each column are in **bold**. † in-domain training data. All methods share the same retriever (E5/Wiki-18), top-$k=3$, training corpus (NQ+HotpotQA), and strict-EM scorer.

With $\ell_t^s = \log \pi_\theta(a_t \mid c_t)$ and $\ell_t^T = \log \pi_{\text{ref}}(a_t \mid c_t^T)$, the OPD loss is

$$L_{\text{OPD}} = -\frac{1}{|M|} \sum_{t \in M} \text{clamp}\big(\ell_t^T - \ell_t^s,\, 0,\, 1\big) \cdot \ell_t^s, \tag{3}$$

where M is the set of response tokens inside a ⟨SEARCH⟩...⟨/SEARCH⟩ span within a failed trajectory. The $\text{clamp}(\cdot)$ coefficient is computed with a stop-gradient on $\ell_t^T - \ell_t^s$, so the only differentiable factor is ℓ_t^s; the update therefore raises ℓ_t^s on tokens where the teacher is more confident than the student. The lower clamp restricts gradient flow to tokens where the hint-conditioned teacher is more confident than the student; the upper clamp prevents outliers from dominating the batch. Answer tokens are left to GRPO. The total objective is

$$L_{\text{total}} = L_{\text{PPO}} + \lambda_{\text{OPD}} \cdot L_{\text{OPD}}, \tag{4}$$

with $\lambda_{\text{OPD}} = 0.01$ in all main runs. Appendix A gives the full algorithm.

4 Experiments

Datasets. We evaluate HindSearch on seven question-answering benchmarks covering both single-hop and multi-hop settings: NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). Training data are the merged NQ and HotpotQA training splits (169k samples); the other five benchmarks are zero-shot. We follow prior work and adopt exact match (EM) as the evaluation metric, with 51,713 validation samples in total, validated every 50 training steps.

Baselines. We compare against a full set of baselines spanning (1) inference without retrieval (Direct, CoT (Wei et al., 2022)); (2) inference with retrieval (RAG (Lewis et al., 2021), IRCoT (Trivedi et al., 2023), Search-o1 (Li et al., 2025)); (3) fine-tuning based methods (SFT, R1-style RL without search (Guo et al., 2025), rejection sampling, and the Search-R1 PPO and GRPO variants (Jin et al., 2025)); and (4) the concurrent simulated-retrieval method ZeroSearch (Sun et al., 2026), whose released checkpoint we evaluate under our retriever. To ensure a fair comparison, all methods use the same backbone, retriever, knowledge corpus, top-k, and training data.

Training details. The backbone is Qwen2.5-3B-Instruct, with E5-base-v2 (Wang et al., 2024) as the retriever over the Wikipedia-18 corpus, top-3 retrieval, and a maximum of $T \leq 4$ turns per trajectory. We use $\pi_\theta = \pi_{\text{ref}} =$ Qwen2.5-3B-Instruct (Qwen et al., 2025) and Qwen3.6-27B (Qwen Team, 2026) as the judge π_J, served via a local API. $\lambda_{\text{OPD}} = 0.01$, GRPO group size $G = 8$, and we train for 300 steps on 2×H200. The OPD pass adds roughly 15% wall-clock per step. Full hyperparameters in Appendix B.

5 Results

Main result. Table 1 reports validation EM averaged across the seven benchmarks. HindSearch exceeds all prior baselines by step 150 and continues climbing throughout training (Figure 2). HindSearch reaches 39.4 at its best checkpoint, ahead

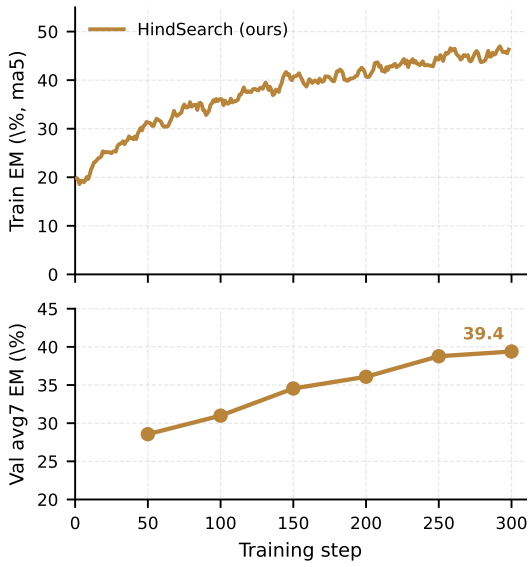

Figure 2: HindSearch training and validation dynamics across 300 GRPO steps. Top: train EM (rolling 5-step mean) climbs steadily to a peak of 49.97. Bottom: val avg EM (dots) rises monotonically from 28.6 at step 50 to 39.4 at step 300.

Configuration	Avg EM
HindSearch (full, 3B-Instruct)	**39.4**
– hindsight (judge w/o $a^\star$)	34.7
PPO-clipped OPD (Wang et al., 2026a)	32.7
Multi-seed (seed 42)	39.2
Per-step earliest-critical	35.5
7B teacher	collapse

Table 2: Ablations on the seven-benchmark search-QA suite (val@300). *7B teacher* collapsed within tens of steps. Additional variants in Appendix C.

of the strongest prior baseline (Search-R1 GRPO, 33.6) by +5.8 points. Training was stable through step 300.

Training curve. Figure 2 shows HindSearch's training and validation dynamics over the 300-step run. Train EM (rolling 5-step mean) climbs steadily to a peak of 49.97, with KL to the anchor staying below 0.43 throughout. Val avg EM rises monotonically from 28.6 at step 50 to 39.4 at step 300, with no plateau visible at the end of training. The OPD term yields a smooth, stable curve under the same outcome reward and GRPO trust region.

Per-task gains. Gains show up on most benchmarks: TriviaQA leads at 60.5, with five others above 30. MuSiQue, the hardest multi-hop benchmark, reaches 14.2—low in absolute terms but in line with prior reports on its difficulty. The improvement transfers cleanly off the training distribution: the two in-domain benchmarks (NQ, HotpotQA) gain +5.85 on average over Search-R1 GRPO, while the five zero-shot benchmarks gain +5.80, indicating that the OPD signal induces general search behavior rather than memorization of training-set surface patterns.

Ablations. Table 2 isolates four design choices. *Hindsight.* Removing the judge's gold-answer access drops the average from 39.4 to 34.7, erasing 4.7 of the 5.8-point gain over Search-R1 GRPO. This is consistent with hindsight, rather than the act of judging alone, being the primary contributor (a no-answer judge also produces less directive critiques, which is partially confounded with the information change). *Loss form.* Swapping the plain-clamp OPD loss for OpenClaw-RL's PPO-clipped, top-K surrogate (Wang et al., 2026a) drops the average to 32.7, below Search-R1 GRPO: the loss form is load-bearing. We caution that this re-implements an OPD loss designed for GUI/tool-use into search-QA, so the −6.7pp gap may include a loss-form × task-domain interaction. *Broadcast vs. per-step hint.* A per-step "earliest critical" variant reaches 35.5, 3.9 points below the full method; broadcasting the critique to every search step is a non-trivial design choice. *Teacher choice.* A same-family 7B teacher destabilizes training within tens of steps (Appendix C).

Sensitivity analyses. *Judge size and family.* Replacing Qwen3.6-27B (Qwen Team, 2026) with same-family Qwen2.5-7B-Instruct (Qwen et al., 2025) drops the average to 34.8 (−4.6 pp); a cross-family DeepSeek-V4-Flash (DeepSeek-AI, 2026) judge reaches 39.6 (+0.2 pp). The method transfers across judge families but degrades when the judge is substantially weaker than the student's critique-following capacity. *Answer leakage.* The judge sees the gold answer, so its critique can surface the answer string. Three facts bound the concern: the judge is absent at inference; the OPD loss is masked to search-action tokens and never updates the final-answer span; and the in-domain (+5.85pp) and zero-shot (+5.80pp) gains are nearly identical, inconsistent with memorization-via-leakage as the dominant mechanism.

Limitations

Backbone and scale. All main experiments use Qwen2.5-3B-Instruct; Table 2 includes a second-seed run. We focus on the instruction-tuned variant: applying HindSearch from a pure base model (Qwen2.5-3B-Base) requires a format-acquisition stage—GRPO and the OPD signal can only act once the policy emits well-formed `<search>`/`<answer>` actions—which is out of scope for this short paper.

Retriever and concurrent baselines. We use E5-base-v2 (Wang et al., 2024) over the Wikipedia-18 corpus, top-$k = 3$, following the standard search-RL evaluation setup (Jin et al., 2025; Chen et al., 2025; Shi et al., 2025). Numbers may shift under sparse retrievers (e.g., BM25). Concurrent work such as ZeroSearch (Sun et al., 2026) targets an orthogonal axis—training-time retrieval cost—under a different retrieval setup; a head-to-head comparison would require matching its retriever and reward, and we do not attempt it here.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. *Preprint*, arXiv:2306.13649.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. 2018. Hindsight experience replay. *Preprint*, arXiv:1707.01495.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025. Research: Learning to reason with search for llms via reinforcement learning. *Preprint*, arXiv:2503.19470.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, and 4 others. 2025. Process reinforcement through implicit rewards. *Preprint*, arXiv:2502.01456.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence. Technical report.

Liang Ding. 2026. Agenther: Hindsight experience replay for llm agent trajectory relabeling. *Preprint*, arXiv:2603.21357.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *Preprint*, arXiv:2011.01060.

Michael Y. Hu, Benjamin Van Durme, Jacob Andreas, and Harsh Jhamtani. 2026. Sample-efficient online learning in lm agents via hindsight trajectory rewriting. *Preprint*, arXiv:2510.10304.

Ziyang Huang, Xiaowei Yuan, Yiming Ju, Jun Zhao, and Kang Liu. 2025. Reinforced internal-external knowledge synergistic reasoning for efficient adaptive search agent. *Preprint*, arXiv:2505.07596.

Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025. s3: You don't need that much data to train a search agent via rl. *Preprint*, arXiv:2505.14146.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Guankai Li, Jiabin Chen, Yi Xu, Xichen Zhang, and Yuan Lu. 2026. Hypereyes: Dual-grained efficiency-aware reinforcement learning for parallel multimodal search agents. *Preprint*, arXiv:2605.07177.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *Preprint*, arXiv:2501.05366.

Zihan Liang, Yufei Ma, Ben Chen, Zhipeng Qian, Huangyu Dai, Lingtao Mao, Xuxin Zhang, Chenyi Lei, and Wenwu Ou. 2026a. Ig-search: Step-level information gain rewards for search-augmented reasoning. *Preprint*, arXiv:2604.15148.

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

Shichao Ma, Zhiyuan Ma, Ming Yang, Xiaofan Li, Xing Wu, Jintao Du, Yu Cheng, Weiqiang Wang, Qiliang Liu, Zhengyang Zhou, and Yang Wang. 2026a. Tspo: Breaking the double homogenization dilemma in multi-turn search policy optimization. *Preprint*, arXiv:2601.22776.

Yufei Ma, Zihan Liang, Ben Chen, Zhipeng Qian, Huangyu Dai, Lingtao Mao, Xuxin Zhang, Chenyi Lei, and Wenwu Ou. 2026b. Sd-search: On-policy hindsight self-distillation for search-augmented reasoning. *Preprint*, arXiv:2605.18299.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Preprint*, arXiv:2212.10511.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. *Preprint*, arXiv:2210.03350.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27b dense model. Hugging Face model card.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

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

Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. 2025. Search and refine during think: Facilitating knowledge refinement for improved retrieval-augmented reasoning. *Preprint*, arXiv:2505.11277.

Jiangming Shu, Yuxiang Zhang, Ye Ma, Xueyuan Lin, and Jitao Sang. 2026. Evaluate-as-action: Self-evaluated process rewards for retrieval-augmented agents. *Preprint*, arXiv:2603.09203.

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

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. 2026. Zerosearch: Incentivize the search capability of llms without searching. *Preprint*, arXiv:2505.04588.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Yinjie Wang, Xuyang Chen, Xiaolong Jin, Mengdi Wang, and Ling Yang. 2026a. Openclaw-rl: Train any agent simply by talking. *Preprint*, arXiv:2603.10165.

Zhao Wang, Ziliang Zhao, and Zhicheng Dou. 2026b. Prorag: Process-supervised reinforcement learning for retrieval-augmented generation. *Preprint*, arXiv:2601.21912.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. 2025. Stepsearch: Igniting llms search ability via stepwise proximal policy optimization. *Preprint*, arXiv:2505.15107.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*.

Tongyu Wen, Guanting Dong, and Zhicheng Dou. 2026. Smartsearch: Process reward-guided query refinement for search agents. *Preprint*, arXiv:2601.04888.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, and 4 others. 2025. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *Preprint*, arXiv:2509.08755.

Tianle Xia, Ming Xu, Lingxiang Hu, Yiding Sun, Wenwei Li, Linfang Shang, Liqun Liu, Peng Shu, Huan Yu, and Jie Jiang. 2026. Search-p1: Path-centric reward shaping for stable and efficient agentic rag training. *Preprint*, arXiv:2602.22576.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Qingfei Zhao, Ruobing Wang, Dingling Xu, Daren Zha, and Limin Liu. 2025. R-search: Empowering llm reasoning with search via multi-reward reinforcement learning. *Preprint*, arXiv:2506.04185.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *Preprint*, arXiv:2504.03160.

A Algorithm pseudocode (extended)

We give the full HindSearch training step, with bookkeeping omitted from the main-paper sketch.

Algorithm 1 HindSearch training step (full)

Require: policy π_θ, frozen teacher π_{ref}, judge π_J, batch $\mathcal{B}$, group size G, weight λ_{OPD}

For each prompt $p \in \mathcal{B}$, rollout G trajectories $\{\tau_g^{(p)}\}_{g=1}^{G}$ with π_θ.
Compute EM rewards $r(\tau_g^{(p)})$ and group-normalized advantages $\hat{A}_g^{(p)}$.
$\mathcal{T}_- \leftarrow \{\tau : r(\tau) = 0\}$.
for $\tau \in \mathcal{T}_-$ in parallel **do**
  $h(\tau) \leftarrow \pi_J(\tau, a_\tau^\star)$ {single sample, temperature 0.7, max 96 tokens}
end for
Build $M \leftarrow \{(\tau, t) : \tau \in \mathcal{T}_-,\ a_t \in \langle\text{SEARCH}\rangle \text{ span}\}$.
For each $(\tau, t) \in M$, build c_t^T via Eq. 2.
Forward π_{ref} on $\{c_t^T\}_{(\tau,t)\in M}$; collect ℓ_t^T.
Forward π_θ on $\mathcal{B}$; collect ℓ_t^s on all response tokens.
$L_{\text{PPO}} \leftarrow$ standard GRPO clipped objective with retrieved-token mask.
$L_{\text{OPD}} \leftarrow$ Eq. 3 on M.
Step optimizer on $L_{\text{PPO}} + \lambda_{\text{OPD}} L_{\text{OPD}}$.

B Hyperparameters

Table 3 gives the full hyperparameter set used in all main and ablation runs.

Setting	Value
Backbone	Qwen2.5-3B-Instruct
Frozen teacher	Qwen2.5-3B-Instruct (init)
Judge model	Qwen3.6-27B (Qwen Team, 2026)
Training data	NQ + HotpotQA (169k)
Retriever	E5-base-v2 (Wikipedia-18)
Top-k retrieval	3
Max turns T	4
RL algorithm	GRPO
Group size G	8
PPO clip ϵ	0.2
KL coef (anchor)	0.001
λ_{OPD}	0.01
OPD token mask	search-span only, $\mathcal{T}_-$ only
OPD clamp range	$[0, 1]$
Optimizer	AdamW
LR	1×10^{-6}
LR schedule	constant
Train steps	300
Validation every	50 steps
Hardware	2×H200
Wall-clock overhead	$\sim$15% vs. unmodified GRPO

Table 3: Full hyperparameters for the main HindSearch runs.

C Negative results

We collect here variants we tried that did not work or underperformed the default HindSearch configuration.

Per-step earliest-critical critique. Instead of broadcasting one trajectory-level critique $h(\tau)$ to every search step, we asked the judge to pick a single "earliest critical" search step and emitted a localized hint only there. This variant reached `val@50` $= 28.55$, comparable to the trajectory-broadcast value at the same step (28.56), but it never closed the gap and ended well below the trajectory-broadcast `val@150` $= 34.55$. Looking at intermediate checkpoints, the localized hint seems too narrow: when the critique is genuinely trajectory-level (e.g., "the agent should have re-decomposed the question after the third turn"), pinning it to one token span fits poorly and the OPD gradient gets noisy.

Same-family 7B teacher. The obvious way to widen the teacher–student capability gap is a strictly larger same-family teacher. We tried Qwen2.5-7B-Instruct as π_{ref} with the 3B student. Training collapsed within `79` steps: validation EM dropped and KL to the anchor diverged. Our reading is that the 7B teacher disagrees with the 3B student on routine search tokens regardless of the hint, and at $\lambda_{\text{OPD}} = 0.01$ this produces a persistent capability-gap gradient that pulls the student off the GRPO trust region rather than nudging it toward repair behaviors. The frozen 3B copy avoids this because its disagreement with the student is mediated entirely by the hint context.

Token-level judging. A variant differing from HindSearch only in the judge output format—per-token categorical labels over the emitted search query instead of a natural-language critique—collapsed within tens of steps. Logs showed the per-token labels were dominated by a single class, producing a degenerate OPD target. We kept natural-language critiques in the main method.

Advantage-injection OPD on EM-rewarded search. We tried a direct port of OpenClaw-RL's advantage-injection formulation (Wang et al., 2026a) to our search setup, sweeping the injection weight and using both rule-based and judge-based hints. All injection variants collapsed within 25–85 steps. Inspecting the runs, on EM-rewarded search the outcome advantage and the directive log-probability gap often disagree in sign on the same token, so a single combined scalar advantage produces updates that neither optimize EM nor follow the directive. The separate-loss form in Eq. 4 sidesteps this by keeping the two signals orthogonal in loss space.

λ-decay schedules. We tested linearly decaying λ_{OPD} from 0.01 to 0 over $[0.2T, 0.4T]$ and over $[0.0T, 0.5T]$, on the intuition that OPD should matter most early in training. Both schedules underperformed the constant $\lambda_{\text{OPD}} = 0.01$ at val@150; we use the constant schedule in all main results.

D Qualitative critique example

Figure 3 shows a representative failed trajectory from training (HotpotQA) together with the trajectory-level critique returned by the judge. Snippets are abbreviated to fit the column. The hint is then prepended to the student's context for every `<search>` step in this trajectory, and the OPD loss pushes the student toward the teacher's hint-conditioned token distribution on those search tokens only.

Question. *Who was the director of the 1985 film Brazil?*
Gold answer. Terry Gilliam.

Trajectory ($r(\tau) = 0$):

`Step 1.` `<search>`*Who was the director of Brazil?*`</search>`
Retrieved: Doc 1 "Cinema of Brazil" (Glauber Rocha); Doc 2 "Brazil (country)" (Fernando Meirelles, *City of God*); Doc 3 "Brazil (1985 film)" (production background; no director name in snippet).

`Step 2.` `<search>`*director of Brazil*`</search>`
Retrieved: Doc 1–2 "Roberto Azevêdo" (WTO Director-General); Doc 3 "Secretary General of Foreign Affairs of Brazil".

`Step 3.` `<answer>`*Glauber Rocha*`</answer>`

Judge critique.

The agent confused the country "Brazil" with the 1985 film of the same name. The first query should have disambiguated to the film (e.g. "director of Brazil 1985 film" or "Terry Gilliam Brazil 1985"); the second query made the ambiguity worse by retrieving political-figure pages instead.

Figure 3: Representative failed trajectory and trajectory-level critique. The same critique is broadcast as a hint to all three `<search>` steps and used to condition the frozen teacher in the OPD loss.

E Judge prompt and model identifiers

Judge model identifiers. For full reproducibility, the exact judge checkpoints used in this work are:

- **Qwen3.6-27B** (main): `Qwen/Qwen3.6-27B` on Hugging Face, served via vLLM with `enable_thinking=false`.

- **Qwen2.5-7B-Instruct** (sensitivity ablation): `Qwen/Qwen2.5-7B-Instruct`, served via vLLM.
- **DeepSeek-V4-Flash** (sensitivity ablation): public DeepSeek API, model identifier `deepseek-v4-flash`, endpoint `https://api.deepseek.com/chat/completions`.

All judge calls use `temperature=0.2`, `max_tokens=500`, and the prompt below.

Trajectory-judge prompt. The frozen judge π_J is queried once per failed trajectory with the following template (the first and last two search steps are kept verbatim; intermediate steps are abbreviated to control prompt length):

```
A search agent failed to
answer the question correctly.
Diagnose the failure.

Question: {question}
Correct answer: {ground_truth}

Agent's search trajectory:
Step 1: query = "..."
  retrieved: ...
... omitted N middle steps ...
Step k: query = "..."
  retrieved: ...

In one or two sentences (max
50 words), describe what
the agent should have done
differently. Be specific to
THIS trajectory -- name actual
queries, retrieved evidence, or
missing information. Output the
critique only -- no preamble, no
labels, no markdown.
```

The judge's response is parsed as plain text, capped at 500 characters by extracting up to the last sentence boundary, and used directly as the hint $h(\tau)$ that conditions the frozen teacher (Eq. 2).

F Training cost vs. canonical OPD

HindSearch's compute overhead is comparable to or lower than canonical on-policy distillation in two respects. First, the auxiliary teacher in canonical OPD is queried once per response token to obtain π_{ref} log-probabilities; HindSearch keeps the same per-token teacher forward but restricts the loss to the search-action token mask M, which in our trajectories covers roughly 20–30% of response tokens, so the OPD gradient computation is correspondingly cheaper than a full-response OPD. Second, the judge call is paid only on failed trajectories ($r(\tau) = 0$, typically 60–80% of rollouts early in training and declining as the policy improves); this adds about 15% wall-clock per step in our setup—a one-time cost at training time that disappears at inference, since the judge is not in the rollout loop at evaluation. Hint-conditioned distillation methods of the same family pay a comparable overhead.