OPDSearch+: On-Policy Distillation with RL Refinement for Search-Augmented Reasoning

Qinglin Ye[1,2,*], Zhiyuan Gu[1,3,*], Jingjie Xia[1,2,*], Yiheng Zhang[4], Kaiyan Zhao[5], Shunchao Zheng[6], Yuhang Mu[7], Wenchao Du[1], Yiming Wang[8,†]

[1]University of Chinese Academy of Sciences [2]Institute of Computing Technology, Chinese Academy of Sciences [3]Institute of Automation, Chinese Academy of Sciences
[4]University of Macau [5]Wuhan University [6]Georgia Institute of Technology [7]Northwestern Polytechnical University
[8]University of Hong Kong
[*]Equal contribution. [†]Corresponding author.

Abstract

Search-augmented reasoning remains difficult for small language models. On-policy distillation (OPD) from trained teachers offers a promising direction, but suffers from two issues: (1) high-quality multi-turn search trajectories depend on dynamic retriever responses, making SFT data prohibitively expensive to collect at scale; (2) task-specifically trained teachers incur substantial training cost, while directly applying OPD with an off-the-shelf teacher without task-specific fine-tuning constrains the student to the teacher's performance ceiling and suffers from severe training instability. We propose **OPDSearch+**, the first distillation paradigm that requires no teacher fine-tuning for search-augmented reasoning. We investigate the role of a frozen off-the-shelf instruct model as the teacher in on-policy distillation, and reveal a key insight: *the teacher reshapes the student's policy distribution so that subsequent RL converges to a superior solution that RL alone cannot reach*. In stage one, the student interacts with a live search engine and is distilled via a per-position forward KL objective, transferring reasoning decomposition and evidence integration skills without any task-specific teacher training. In stage two, RL refines the distilled student from a richer behavioral foundation, achieving performance that RL alone cannot reach from scratch. Across seven QA benchmarks, OPDSearch+ with a 3B model consistently outperforms all prior 3B RL baselines, achieving gains of 13.1% on HotpotQA and 8.5% on 2WikiMultihopQA.

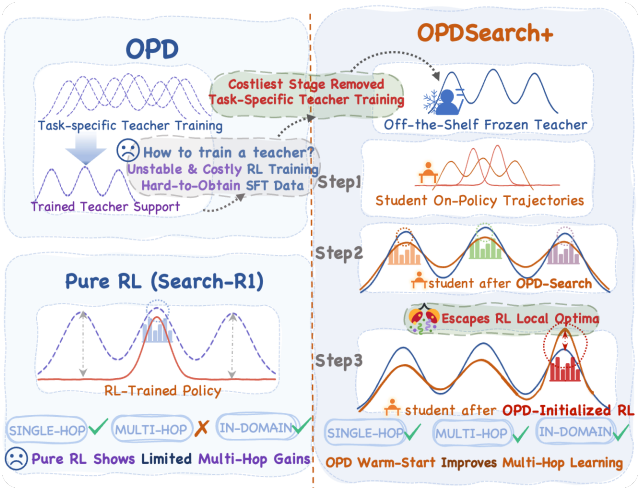

Figure 1: Left: Prior OPD requires costly task-specific teacher training; pure RL shows limited multi-hop gains. Right: OPDSearch+ eliminates teacher training by using a frozen off-the-shelf teacher. The teacher reshapes the student's distribution via on-policy distillation, enabling subsequent RL to escape local optima and achieve strong single-hop, multi-hop, and in-domain performance.

1 Introduction

Training LLMs to interact with search engines for knowledge-intensive QA has become a central research direction (Schick et al. 2023; Yao et al. 2023; Nakano et al. 2021). On-policy distillation (OPD) from trained teachers offers a promising paradigm for transferring search-and-reason capabilities from larger models to smaller ones. However, existing OPD methods face two fundamental obstacles. First, high-quality multi-turn search trajectories depend on dynamic retriever responses, making SFT data prohibitively expensive to collect at scale, and unlike static text generation, these trajectories cannot be reused across different retrieval systems or corpus versions. Second, task-specifically trained teachers incur substantial training cost and instability, while directly applying OPD with an off-the-shelf teacher without task-specific fine-tuning constrains the student to the teacher's performance ceiling and suffers from severe training instability due to distributional misalignment between the teacher and the student's on-policy trajectories.

We propose **OPDSearch+**, a distillation paradigm that requires *no teacher fine-tuning* and resolves both obstacles. Our key insight is that *the teacher's role is not to provide a performance ceiling for imitation, but to reshape the student's policy distribution so that subsequent RL converges to a superior solution that RL alone cannot reach*. Concretely, a frozen off-the-shelf instruct model serves as the teacher: it provides token-level supervision via a per-position forward KL objective on student-generated trajectories, transferring reasoning decomposition and evidence integration capabilities without any task-specific teacher training. The student interacts with a live search engine, and the teacher evaluates these on-policy trajectories, avoiding both the data con-

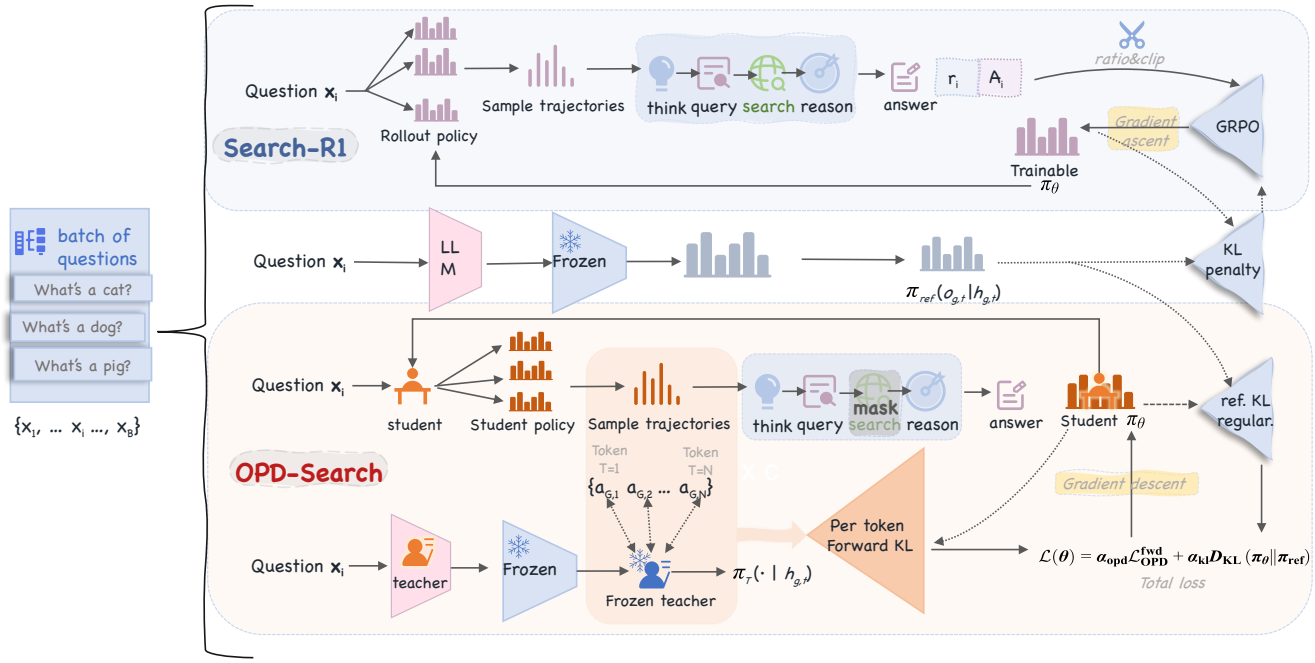

Figure 2: Comparison between RL-based search agent training (Search-R1) and our On-Policy Distillation (OPDSearch+) framework. (a) Search-R1 uses outcome reward (EM/F1) to train the model via GRPO, which conflates retrieval quality with answer correctness and suffers from reward hacking. (b) OPDSearch+ generates on-policy trajectories from the student interacting with a live search engine, then distills the teacher's token-level distribution onto these trajectories. The teacher's search behavior serves as implicit supervision for both reasoning and retrieval, providing a strong initialization for subsequent RL refinement.

struction challenge and the teacher training cost. Once distillation has expanded the student's expressive capacity, RL refines the distilled student from this richer behavioral foundation, achieving performance that RL alone cannot reach from scratch.

OPDSearch+ achieves mean EM 0.4402 across seven QA benchmarks, exceeding all 3B RL baselines including the previous best GiGPO-Instruct at 0.421. Gains are particularly pronounced on multi-hop tasks: $+13.1\%$ on HotpotQA and $+8.5\%$ on 2Wiki over the best 3B baseline AutoRefine-Base.

Our contributions are listed as follows:

- We propose OPDSearch+, the first on-policy distillation framework for interactive retrieval that uses a frozen off-the-shelf teacher, providing token-level supervision on live search trajectories without costly data construction or task-specific teacher training.
- We show that OPD reshapes the student policy into a stronger initialization for RL, enabling it to outperform both the teacher and RL from the same base model. Theoretically, we establish that clipped forward KL controls gradient variance under teacher–student mismatch; empirically, it remains more stable than reverse-KL variants while increasing policy entropy, broadening behavioral diversity, and reducing search turns.
- The resulting two-stage pipeline achieves mean EM 0.4402 across seven QA benchmarks with a 3B student, outperforming all 3B baselines (the best prior result is 0.421). Multi-hop gains are particularly pronounced: $+13.1\%$ on HotpotQA and $+8.5\%$ on 2WikiMultiHopQA, demonstrating that distillation transfers compositional reasoning more efficiently than RL discovers it from scratch.

2 Related Work

RL for Search-Augmented Reasoning. Search-R1 (Jin et al. 2025) demonstrates that LLMs can learn to interleave reasoning with retrieval via RL with outcome reward. Subsequent work including R1-Searcher (Song et al. 2025), ReSearch (Chen et al. 2025), and ASearcher (Gao et al. 2025) extends this to diverse settings but retains the outcome-only reward paradigm. Process-reward approaches such as StepSearch (Wang et al. 2025) and GiGPO (Feng et al. 2025) add step-level supervision but require additional engineering (GPT-generated sub-questions, contrastive group construction). All RL-based methods share fundamental challenges with sample efficiency and training stability, particularly for smaller models.

Knowledge Distillation for Reasoning LLMs. Knowledgedistillation (Hinton, Vinyals, and Dean 2015) has been widely applied to compress LLMs while preserving reasoning capabilities. Offline approaches (Kim and Rush 2016) train on teacher-generated data but suffer from train-test

distribution mismatch: the student sees teacher trajectories during training but must follow its own distribution at inference, leading to error accumulation. On-policy distillation (OPD) (Agarwal et al. 2024) addresses this by scoring student-generated samples with the teacher, typically using reverse KL minimization. Recent work explores the interplay between KL direction and training stability: EOPD (Jin et al. 2026) uses entropy-adaptive KL mixing, Decoupled-KL (Zhao et al. 2026) analyzes the design space of prefix source and KL direction and finds that forward KL is essential for preventing entropy collapse in long-sequence distillation, and KDRL (Xu et al. 2025) jointly optimizes KL distillation with RL objectives. Recently, several works have begun extending distillation to interactive environments: SD-Search (Ma et al. 2026) performs on-policy self-distillation within a search-augmented reasoning loop via JSD minimization at search-query positions (using the same model as both teacher and student); Agent Distillation (Kang et al. 2025) distills full agent trajectories including retrieval and code tool calls; TT-OPD (Jeong 2026) applies turn-level truncated OPD in multi-tool interactive settings. However, these methods either rely on self-distillation (unable to leverage stronger models' capabilities), require task-specifically trained teachers, or use mode-seeking objectives (JSD/reverse KL). In contrast, OPDSearch+ uses a **frozen off-the-shelf instruct model** (requiring no task-specific fine-tuning) as a cross-model teacher and employs a **per-position forward KL** objective (a clipped importance-weighted estimator on student prefixes) for distillation in an interactive retrieval environment—this combination encourages the student to both leverage the large model's generalization capabilities and preserve diversity of search strategies, without any expensive teacher training pipeline.

KD-RL Hybrid Methods. Recent methods have explored combining KD with RL for reasoning tasks: RLAD (Zhang et al. 2026) uses advantage-weighted trust-region distillation, SPOT (Lin and Han 2026) uses proximal on-policy distillation as RL initialization, TGPO (Liu et al. 2026) has the teacher generate in the student's context, and SC-GRPO (Shan et al. 2026) uses self-conditioned KL as credit-assignment weights. These methods all operate in static text-generation settings (e.g., mathematical reasoning) where the model does not interact with external systems. In contrast, OPDSearch+ addresses the distinct challenge of *interactive retrieval environments*, where trajectories depend on dynamic search engine responses and the teacher must provide guidance conditioned on live environment feedback.

3 Method

3.1 Why On-Policy Distillation Suits Search Agents

We establish three formal properties that justify why forward-KL on-policy distillation is particularly well-suited to search-augmented reasoning agents.

Proposition 1 (Implicit multi-level supervision)**.** *Let a trajectory $o = (o_1, \ldots, o_T)$ be partitioned into reasoning tokens $\mathcal{T}_R$, query tokens $\mathcal{T}_Q$, and answer tokens $\mathcal{T}_A$ with $\mathcal{T}_R \cup \mathcal{T}_Q \cup \mathcal{T}_A = \mathcal{M}$. Define the teacher-to-student importance ratio $r_t = \pi_{tea}(o_t \mid o_{<t})/\pi_\theta(o_t \mid o_{<t})$ and its clipped version $\bar{r}_t = \mathrm{clip}(r_t, \epsilon, R_{max})$. Then the implemented forward-KL gradient decomposes as:*

$$\nabla_\theta \mathcal{L}_{OPD}^{fwd} = -\frac{1}{|\mathcal{M}|}\left(\sum_{t\in\mathcal{T}_R} \bar{r}_t\, g_t + \sum_{t\in\mathcal{T}_Q} \bar{r}_t\, g_t + \sum_{t\in\mathcal{T}_A} \bar{r}_t\, g_t\right), \tag{1}$$

where $g_t = \nabla_\theta \log \pi_\theta(o_t \mid o_{<t})$. Consequently, for query tokens where the teacher assigns high probability but the student does not, a raw ratio $r_t > 1$ amplifies the gradient through $\bar{r}_t$, up to the cap R_{max}—providing implicit query-quality supervision without a dedicated retrieval reward.

The decomposition follows from linearity of summation over token positions. The key consequence is that the clipped ratio $\bar{r}_t$ acts as a *per-token adaptive reward*: tokens the teacher strongly endorses receive amplified gradients, up to $R_{\max}$, regardless of their functional role in the trajectory. A well-formed entity-specific query can receive a large raw ratio r_t when the teacher favors it but the student under-assigns probability, while a verbatim question-copy receives $r_t \approx 1$. This provides implicit reward shaping that outcome-based RL cannot achieve—outcome rewards assign identical credit to all tokens in the trajectory.

Proposition 2 (Second-moment control under distributional mismatch)**.** *Let $g_t = \nabla_\theta \log \pi_\theta(o_t \mid o_{<t})$ and let $\bar{r}_t = \mathrm{clip}(r_t, \epsilon, R_{max})$ be the detached clipped importance ratio. Then the per-token gradient second moment under the implemented forward-KL objective satisfies:*

$$\mathbb{E}_{o_t\sim\pi_\theta}\left[\|\bar{r}_t\, g_t\|^2\right] \;\le\; R_{max}^2 \cdot \mathbb{E}_{o_t\sim\pi_\theta}\left[\|g_t\|^2\right]. \tag{2}$$

In contrast, for the reverse-KL surrogate, the effective weight $c_t = \pi_\theta(o_t)/\pi_{tea}(o_t) - 1$ is unbounded and correlates positively with the sampling probability $\pi_\theta(o_t)$, so high-variance gradient terms are encountered frequently rather than rarely.

Since $\bar{r}_t \in [\epsilon, R_{\max}]$ and is treated as stop-gradient, the bound follows from $\|\bar{r}_t\, g_t\|^2 \le R_{\max}^2 \|g_t\|^2$ pointwise. The critical asymmetry is: under forward-KL, tokens with a large raw ratio r_t receive a capped coefficient $\bar{r}_t \le R_{\max}$ and are those where π_θ is small—precisely the tokens *rarely sampled* on-policy, so their high-magnitude contributions appear infrequently in minibatches. Under reverse-KL, the coefficient $c_t = \pi_\theta(o_t)/\pi_{\mathrm{tea}}(o_t) - 1$ grows large when the student over-assigns probability relative to the teacher, and these tokens are sampled *frequently* (proportional to π_θ), creating a positive feedback loop that drives entropy collapse. This explains the empirical instability of reverse-KL variants observed in Section 4.4.

Proposition 3 (Distribution-shift bound for on-policy evaluation)**.** *Let $P_\theta^{\mathcal{R}}$ denote the prefix distribution induced by the student interacting with retriever $\mathcal{R}$, and $P_{offline}$ the distribution of pre-collected offline trajectories. Define the per-prefix forward KL $f(o_{<t}) = D_{KL}\big(\pi_{tea}(\cdot \mid o_{<t}) \,\|\, \pi_\theta(\cdot \mid o_{<t})\big)$ bounded by C. Then:*

$$\left|\mathbb{E}_{P_\theta^{\mathcal{R}}}[f] - \mathbb{E}_{P_{offline}}[f]\right| \;\le\; C \cdot D_{TV}\big(P_\theta^{\mathcal{R}},\, P_{offline}\big). \tag{3}$$

On-policy distillation sets $P_{offline} = P_\theta^{\mathcal{R}}$, eliminating the distribution-shift gap entirely.

The bound follows from treating the per-prefix KL as a bounded test function and applying the variational characterization of total variation distance. In search-augmented settings, the prefix includes retrieval results that depend on prior queries issued by π_θ: even small policy changes alter which passages are retrieved, cascading through subsequent reasoning steps. Offline methods, which train on teacher-generated trajectories, suffer from this compounding distribution shift—the student sees teacher retrieval contexts during training but must follow its own at inference. On-policy distillation avoids this entirely by evaluating the teacher on the student's *actual* retrieval contexts, ensuring the gradient always reflects the student's real operating conditions.

3.2 Problem Setup

We consider the search-augmented QA task following the Search-R1 framework (Jin et al. 2025). Given a question x, the model generates a multi-turn trajectory consisting of interleaved reasoning (`<think>`), search queries (`<search>`), retrieved passages (`<information>`), and a final answer (`<answer>`). The search engine $\mathcal{R}$ returns relevant passages for each query. A trajectory $o = (o_1, o_2, \ldots, o_T)$ is the full sequence of tokens generated by the model (reasoning, queries, answers), where retrieved passages are provided by the environment and not generated by the model.

3.3 On-Policy Distillation for Search (OPDSearch+)

Overview. OPDSearch+ trains a student policy π_θ (Qwen2.5-3B) to imitate a frozen teacher π_{tea} (Qwen2.5-14B-Instruct) *on the student's own search trajectories*. The key distinction from standard text-only OPD is that trajectories are generated through interaction with a live retrieval environment: the student issues search queries, receives real passages from the Wikipedia corpus, and must integrate this evidence. The teacher evaluates the student's complete trajectory (including the environmental feedback) and provides token-level supervision.

Training procedure. For each training batch of questions $\{x_1, \ldots, x_B\}$:

1. **On-policy rollout**: The student π_θ generates G trajectories per question by interacting with the search engine $\mathcal{R}$. Each trajectory $o_i^{(g)}$ includes the student's reasoning, queries, and the search engine's responses.
2. **Teacher scoring**: The frozen teacher π_{tea} computes token-level log-probabilities on the student-generated trajectories (excluding retrieved-passage tokens, which are environment-provided).
3. **Student update**: The student parameters are updated to minimize the KL divergence between the student and teacher distributions on these trajectories.

Loss function. The training objective combines two terms:

$$\mathcal{L}(\theta) = \alpha_{\text{opd}} \cdot \mathcal{L}_{\text{OPD}}(\theta) + \alpha_{\text{kl}} \cdot \mathcal{L}_{\text{KL}}(\theta), \tag{4}$$

where $\mathcal{L}_{\text{OPD}}$ is the forward-KL distillation loss (a clipped importance-weighted estimator; see below) and $\mathcal{L}_{\text{KL}}$ is a KL regularization term against a frozen reference policy (the initial student checkpoint) to prevent catastrophic deviation.

Teacher-weighted cross-entropy (forward KL motivation). Our distillation objective minimizes the *per-position forward KL* $D_{\text{KL}}(\pi_{\text{tea}} \| \pi_\theta)$ at each token position, estimated via importance weighting on student-generated prefixes. Let $\ell_t = \log \pi_\theta(o_t | o_{<t}, x; \mathcal{R})$ and $\ell_t^{\text{tea}} = \log \pi_{\text{tea}}(o_t | o_{<t}, x; \mathcal{R})$, and define the teacher-to-student ratio and its clipped version as $r_t = \exp(\ell_t^{\text{tea}} - \ell_t)$ and $\bar{r}_t = \text{clip}(r_t, \epsilon, R_{\text{max}})$, respectively. The implemented loss is:

$$\mathcal{L}_{\text{OPD}}^{\text{fwd}}(\theta) = -\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \text{sg}(\bar{r}_t) \cdot \log \pi_\theta(o_t | o_{<t}, x; \mathcal{R}), \tag{5}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $\mathcal{M}$ is the set of model-generated token positions (excluding retrieved passages via state masking). When $\bar{r}_t = r_t$ (no clipping), the expectation under $o_t \sim \pi_\theta$ is gradient-equivalent to minimizing $D_{\text{KL}}(\pi_{\text{tea}}(\cdot|o_{<t}) \| \pi_\theta(\cdot|o_{<t}))$ averaged over student-sampled prefixes. In implementation, ϵ=10^{-6} and R_{max}=10; clipping introduces controlled bias relative to the exact forward-KL gradient while limiting importance-weight magnitude. The implemented gradient is:

$$\nabla_\theta \mathcal{L}_{\text{OPD}}^{\text{fwd}} = -\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \text{sg}(\bar{r}_t) \cdot \nabla_\theta \log \pi_\theta(o_t), \tag{6}$$

a weighted policy gradient where tokens with $r_t > 1$ (teacher favors more than student) receive amplified gradients up to the cap R_{max}, encouraging the student to cover the teacher's distribution.

Reverse KL surrogate variant. We also experiment with a convex surrogate for the reverse KL objective $D_{\text{KL}}(\pi_\theta \| \pi_{\text{tea}})$, using $f(u) = e^u - u - 1$ applied to $u_t = \ell_t - \ell_t^{\text{tea}}$:

$$\mathcal{L}_{\text{OPD}}^{\text{rev}}(\theta) = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left[\exp\left(\ell_t - \ell_t^{\text{tea}}\right) - \left(\ell_t - \ell_t^{\text{tea}}\right) - 1 \right], \tag{7}$$

where ℓ_t^{tea} is detached. Unlike forward-KL where high-weight tokens are rarely sampled, this surrogate's high-coefficient tokens have high π_θ and are frequently sampled, making gradients prone to entropy collapse (Appendix G of the supplementary material).

Regularization and masking. A KL penalty $\mathcal{L}_{\text{KL}} = D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ against the frozen initial checkpoint prevents catastrophic deviation. State masking excludes retrieved-passage tokens from all losses.

3.4 Connection to RL and Objective Choice

OPDSearch+ can be viewed as a form of policy gradient with $R_t = \text{sg}(\bar{r}_t)$ as a per-token reward derived from the teacher, providing dense, stable supervision at every token (unlike trajectory-level outcome rewards). The trade-off is that this signal reflects teacher preferences rather than task-specific outcomes, motivating the subsequent RL stage. A detailed comparison of forward-KL vs. reverse-KL gradient properties is provided in Appendix G of the supplementary material.

Method	Size	Single-Hop QA			SH-Avg	Multi-Hop QA				Avg
		NQ†	TriviaQA⋆	PopQA⋆		HotpotQA†	2Wiki⋆	Musique⋆	Bamboogle⋆	
3B Methods (same model size as OPDSearch+)										
Direct Generation◇	3B	0.106	0.288	0.108	0.167	0.149	0.244	0.020	0.024	0.134
SFT◇	3B	0.249	0.292	0.104	0.215	0.186	0.248	0.044	0.112	0.176
Naive RAG◇	3B	0.348	0.544	0.387	0.426	0.255	0.226	0.047	0.080	0.270
Search-o1◇	3B	0.238	0.472	0.262	0.324	0.221	0.218	0.054	0.320	0.255
Search-R1-Base◇	3B	0.421	0.583	0.413	0.472	0.297	0.274	0.066	0.128	0.312
Search-R1-Instruct◇	3B	0.397	0.565	0.391	0.451	0.331	0.310	0.124	0.232	0.336
ReSearch-Base◇	3B	0.427	0.597	0.430	0.485	0.305	0.272	0.074	0.128	0.319
ReSearch-Instruct◇	3B	0.365	0.571	0.395	0.444	0.351	0.272	0.095	0.266	0.331
AutoRefine-Base◇	3B	0.467	0.620	0.450	0.512	0.405	0.393	0.157	0.344	0.405
AutoRefine-Instruct◇	3B	0.436	0.597	0.447	0.493	0.404	0.380	0.169	0.336	0.396
StepSearch-Base◇	3B	—	—	—	—	0.329	0.339	0.181	0.328	—
StepSearch-Instruct◇	3B	—	—	—	—	0.345	0.320	0.174	0.344	—
GiGPO-Instruct◇	3B	0.420	0.595	0.424	0.480	0.369	0.370	0.126	**0.641**	0.421
Our Method (3B student, 14B-Instruct teacher, OPD → RL)										
OPDSearch+ (Ours)	**3B**	**0.4852**	**0.6226**	**0.4513**	**0.520**	**0.4580**	**0.4264**	**0.1819**	0.4560	**0.4402**

Table 1: Performance comparison on QA benchmarks (Exact Match). OPDSearch+ uses a **3B** student model. † indicates in-distribution training data. ⋆ indicates out-of-distribution test sets. ◇ indicates numbers cited from prior work. Best in **bold**, second best underlined.

4 Experiments

4.1 Experimental Setup

Models and data. The student is **Qwen2.5-3B** (base); the teacher is **Qwen2.5-14B-Instruct** (frozen). Training uses NQ (Kwiatkowski et al. 2019) + HotpotQA (Yang et al. 2018) train splits (∼170k QA pairs). Evaluation covers seven benchmarks: NQ, TriviaQA (Joshi et al. 2017), PopQA (Mallen et al. 2023) (single-hop) and HotpotQA, 2WikiMultihopQA (Ho et al. 2020), MuSiQue (Trivedi et al. 2022), Bamboogle (Press et al. 2023) (multi-hop). Retrieval uses E5-base-v2 (Wang et al. 2024) over the December 2018 Wikipedia dump (Karpukhin et al. 2020) (top-3 passages), identical to Search-R1.

Training configuration. Both stages use the veRL framework (Sheng et al. 2024) on 4× H200 GPUs with lr 1×10^{-6}, G=8 rollouts, $T_{\max}$=4 search turns. *OPD stage*: trained to the planned horizon (over 420 updates), batch 512, α_{opd}=1.0, forward-KL loss, and no reward signal. *RL stage*: initialized from the best OPD checkpoint (step 150) and continued until training collapse was observed, at which point the run was stopped; batch 1024, GRPO with $R = 0.9R_{\text{EM}} + 0.1R_{\text{format}}$, and no teacher. Full hyperparameters in Appendix A of the supplementary material.

Baselines. We compare against Search-R1 (Jin et al. 2025), ReSearch (Chen et al. 2025), AutoRefine (Shi et al. 2025), StepSearch (Wang et al. 2025), GiGPO (Feng et al. 2025), and non-RL baselines (Direct Generation, SFT, Naive RAG, Search-o1). All use 3B models.

4.2 Main Results

Table 1 presents the main results. OPDSearch+ with a **3B model** achieves a mean EM of $\mathbf{0.4402}$, the best among all 3B methods (surpassing GiGPO-Instruct at 0.421).

Single-hop QA. OPDSearch+ achieves NQ 0.4852, TriviaQA 0.6226, PopQA 0.4513 (SH-Avg 0.520), exceeding AutoRefine-Base (0.512), the best 3B method.

Multi-hop QA. The largest gains appear on multi-hop benchmarks: HotpotQA 0.4580 ($+13.1\%$ over AutoRefine-Base), 2Wiki 0.4264 ($+8.5\%$ over AutoRefine-Base), MuSiQue 0.1819 (best among all 3B methods), Bamboogle 0.4560 (second only to GiGPO's 0.641 on this 125-sample set).

Key observation: Distillation excels on multi-hop. We attribute the disproportionate multi-hop gains to the teacher's multi-step decomposition capabilities: the 14B-Instruct teacher naturally generates sub-queries for complex questions, and on-policy distillation transfers these behaviors efficiently. In contrast, 3B RL models must discover such strategies from scratch through exploration with sparse rewards.

4.3 Training Dynamics Analysis

Figure 3 shows the training dynamics. The distillation loss decreases from ∼0.66 to ∼0.20, reflecting an emergent curriculum where supervision intensity naturally decreases as the student improves. Query formulation quality improves progressively—by step 200, queries become entity-specific and decomposed rather than verbatim question copies. Training is stable throughout 501 steps without oscillations or divergence, consistent with the variance analysis in Section 3.4.

4.4 Teacher-Weighted CE vs. Reverse-KL Surrogate

We compare our forward-KL objective against four reverse-KL-motivated alternatives under identical settings (same

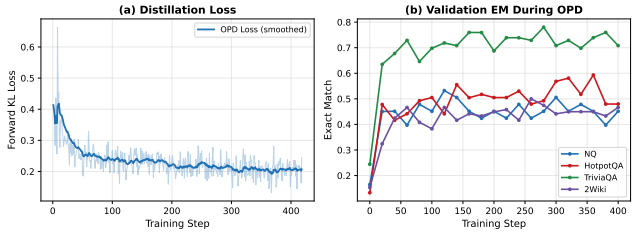

Figure 3: OPD training dynamics (14B→3B, forward-KL). (a) Distillation loss decreases rapidly in the first ∼50 steps and stabilizes around 0.2. (b) Validation EM improves progressively across all benchmarks.

Objective	Stability	Best Val EM	Grad Norm
Forward-KL (Ours)	Stable (400+ steps)	**0.493**	0.73–1.19
Rev-KL Adaptive	Stable (160+ steps)	0.470	0.23–0.32
Rev-KL Clipped	Degrades by step ∼60	0.136	2.2–4.5
Rev-KL JSD	Collapses at step ∼20	—	75.0 → NaN
Rev-KL Log-Ratio	Collapses at step ∼20	—	→ NaN

Table 2: Comparison of distillation objectives under matched hyperparameters (all 3B student, 14B teacher, same lr/α_{kl}/rollout config). Under these shared settings, only the entropy-adaptive variant avoids collapse among reverse-KL alternatives, but it still underperforms our forward-KL objective. Collapses may be mitigable with objective-specific tuning (see text).

teacher, student, lr, KL coefficient; only the loss differs). Results are in Table 2 and Figure 4.

Under shared hyperparameters, JSD and Log-Ratio produce NaN gradients within 20 steps, Clipped degrades by step 100, and only the entropy-adaptive variant (Jin et al. 2026) avoids collapse. Even when stabilized, the adaptive variant achieves lower EM (0.470 vs. 0.493) and saturates by step ∼80 while forward-KL continues improving through step 400. This confirms that reverse-KL objectives require substantially more engineering effort and still underperform in the cross-model search setting, consistent with the gradient variance asymmetry analyzed in Appendix G of the supplementary material.

4.5 Reward Granularity Ablation

We conduct a controlled ablation comparing five RL reward strategies (Pure RL, DAPO token-level, PRIME implicit, Dr. GRPO, Turn-level) against OPD (the reward-granularity figure and Appendix H of the supplementary material). Key findings: (1) OPD never collapses (420+ steps), consistent with the clipped second-moment control of Property 2. (2) All outcome-based RL methods eventually collapse. (3) Dense token-level rewards (DAPO, PRIME) are stable but plateau below OPD→RL's peak. (4) Coarse reward shaping (Dr. GRPO, Turn-level) fails catastrophically within 40–70 steps.

4.6 How OPD Reshapes the Policy Distribution

A central claim of this work is that OPD *reshapes* the student's policy distribution, enabling subsequent RL to converge to solutions unreachable from the base initialization. We provide quantitative evidence (detailed plots in Appendix I of the supplementary material):

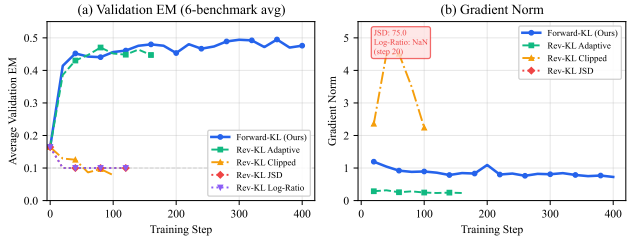

Figure 4: Distillation objective comparison under matched hyperparameters (all 14B→3B). (a) Validation EM: Forward-KL (ours) achieves the highest performance and continues improving through 400 steps. Rev-KL Adaptive avoids collapse but plateaus below forward-KL. Three other reverse variants collapse under default settings. (b) Gradient norm: Forward-KL maintains stable norms (∼0.8–1.2); Adaptive has lower norms (∼0.3) but this does not translate to better EM.

OPD expands behavioral diversity. The OPD-initialized student enters RL with entropy 1.35—significantly higher than the base model's 0.82. OPD does not merely sharpen toward the teacher's preferred actions, but *broadens* the student's action space by transferring diverse search strategies.

Controlled policy drift and search behavior. Forward-KL OPD produces gradual, controlled drift (KL ∼0.2 from reference), while pure RL drifts erratically (KL >0.8 before collapse). OPD also reduces average search turns from 3.5 to 2.7, indicating more precise query formulation that outcome-only RL struggles to discover from sparse rewards.

Gradient stability. Forward-KL OPD maintains monotonically decreasing gradient norms (1.1 → 0.7 over 418 steps), enabling sustained learning. The RL-post-OPD stage starts with low norms (∼0.3), consistent with the OPD-initialized policy being closer to a good solution.

5 Discussion

Off-the-shelf instruct models as teachers. General-purpose instruct models serve as effective OPD teachers because they have internalized information-seeking behaviors through broad instruction-tuning, and the forward-KL objective efficiently extracts these capabilities into the student's on-policy context.

Complementarity with RL. OPD and RL contribute complementary capabilities: OPD establishes search behaviors while RL refines answer accuracy. Joint optimization (0.3660) substantially underperforms the sequential pipeline (0.4402) due to objective interference. The total cost is comparable to Search-R1's RL training (∼48 GPU-hours on 4× H200).

Method	Single-Hop QA			SH-Avg	Multi-Hop QA				Avg
	NQ	TriviaQA	PopQA		HotpotQA	2Wiki	Musique	Bamboogle	
Teacher model evaluated directly (no training)									
Qwen2.5-7B-Instruct (teacher)	0.3019	0.5633	0.3404	0.4019	0.2811	0.2624	0.1059	0.3040	0.3084
Qwen2.5-14B-Instruct (teacher)	0.3630	**0.6479**	0.3919	0.4676	0.3916	0.3513	0.1592	**0.4844**	0.3985
Pure RL baseline (GRPO, no distillation)									
Search-R1-Base (3B)$^{\diamond}$	0.421	0.583	0.413	0.472	0.297	0.274	0.066	0.128	0.312
Search-R1-Instruct (3B)$^{\diamond}$	0.397	0.565	0.391	0.451	0.331	0.310	0.124	0.232	0.336
Pure RL (reproduced, 3B)‡	0.4660	0.6211	**0.4540**	0.5137	0.3467	0.3142	0.0885	0.1760	0.3524
Offline SFT (teacher-generated correct trajectories)									
Offline SFT (14B $\rightarrow$ 3B)	0.3613	0.5315	0.3540	0.4156	0.2945	0.3123	0.1133	0.3040	0.3264
Offline SFT $\rightarrow$ RL (14B $\rightarrow$ 3B)	0.4805	0.6049	0.4293	0.5049	0.4393	0.3887	0.1788	0.4080	0.4185
Pure OPD (forward-KL distillation only, no RL)									
OPD only (14B $\rightarrow$ 3B)	0.3574	0.5872	0.3838	0.4428	0.3602	0.3281	0.1419	0.4000	0.3655
OPD only (7B $\rightarrow$ 3B)	0.3306	0.5534	0.3518	0.4119	0.2872	0.2875	0.1150	0.3920	0.3311
Joint: RL + OPD simultaneously (GRPO + forward-KL, opd_coef=1.0)									
Joint RL+OPD (14B $\rightarrow$ 3B)	0.3597	0.5908	0.3810	0.4438	0.3736	0.3426	0.1463	0.3680	0.3660
Two-stage: OPD warmup $\rightarrow$ GRPO RL fine-tuning									
OPD $\rightarrow$ RL (7B $\rightarrow$ 3B)	0.4665	0.6057	0.4264	0.4995	0.4357	0.3944	**0.1910**	0.4240	0.4205
OPD $\rightarrow$ RL (14B $\rightarrow$ 3B)	**0.4852**	0.6226	0.4513	**0.5197**	**0.4580**	**0.4264**	0.1819	0.4560	**0.4402**

Table 3: Ablation study on teacher scale and two-stage training (full test set evaluation, Exact Match). All methods use Qwen2.5-3B as the student/policy model. "OPD only" denotes pure on-policy distillation without RL. "OPD $\rightarrow$ RL" denotes the two-stage approach (OPD warmup followed by GRPO fine-tuning). "Offline SFT" trains on teacher-generated correct trajectories (EM-filtered). $^{\diamond}$ indicates numbers cited from prior work. ‡ indicates our reproduced RL baseline using $R = 0.9 \cdot R_{\text{EM}} + 0.1 \cdot R_{\text{format}}$ (same reward as our OPD$\rightarrow$RL stage). Best in **bold**.

5.1 Ablation: Teacher Scale and Two-Stage Training

Table 3 presents the ablation on teacher scale and two-stage training.

Pure RL is weak on multi-hop. Our reproduced pure RL baseline achieves competitive single-hop performance (SH-Avg 0.514) but dramatically underperforms on multi-hop: HotpotQA 0.347, 2Wiki 0.314, Bamboogle 0.176 (Avg 0.352). OPD$\rightarrow$RL achieves HotpotQA 0.458 ($+32.1\%$), 2Wiki 0.426 ($+35.7\%$), confirming that OPD provides the multi-step reasoning foundation that pure RL cannot bootstrap from sparse rewards.

OPD outperforms offline SFT. Pure OPD (0.3655) outperforms Offline SFT (0.3264, $+12.0\%$) without trajectory filtering or data construction. Even with RL, Offline SFT$\rightarrow$RL (0.4185) underperforms OPD$\rightarrow$RL (0.4402, $+5.2\%$), supporting Proposition 3 on distribution shift.

Two-stage training is key. OPD$\rightarrow$RL (0.4402) outperforms pure OPD (0.3655, $+20.4\%$), confirming that OPD provides the behavioral foundation while RL refines answer accuracy. Even the 7B-teacher two-stage pipeline (0.4205) exceeds pure OPD with the larger 14B teacher. The corresponding training curve is provided in the supplementary material.

6 Conclusion

We present OPDSearch+, an on-policy distillation framework for search-augmented reasoning that uses a frozen off-the-shelf teacher and requires no task-specific training. The two-stage pipeline (OPD $\rightarrow$ RL) with a 3B student achieves mean EM 0.4402, outperforming all 3B baselines with particularly strong multi-hop gains.

directly applying OPD with an off-the-shelf teacher without task-specific fine-tuning

References

Agarwal, R.; Vieillard, N.; Zhou, Y.; Stanczyk, P.; Ramos, S.; Geist, M.; and Bachem, O. 2024. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*.

Chen, M.; Sun, L.; Li, T.; Sun, H.; Zhou, Y.; Zhu, C.; Wang, H.; Pan, J. Z.; Zhang, W.; Chen, H.; Yang, F.; Zhou, Z.; and Chen, W. 2025. ReSearch: Learning to Reason with Search for LLMs via Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Feng, L.; Xue, Z.; Liu, T.; and An, B. 2025. GiGPO: Group-in-Group Policy Optimization for LLM Agent Training. *arXiv preprint arXiv:2505.10978*.

Gao, J.; Fu, W.; Xie, M.; Xu, S.; He, C.; Mei, Z.; Zhu, B.; and Wu, Y. 2025. Beyond Ten Turns: Unlocking Long-Horizon

Agentic Search with Large-Scale Asynchronous RL. *arXiv preprint arXiv:2508.07976*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Ho, X.; Nguyen, A.-K. D.; Sugawara, S.; and Aizawa, A. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6609–6625.

Jeong, M. 2026. Healthcare AI GYM for Medical Agents. *arXiv preprint arXiv:2605.02943*.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arık, S. Ö.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. In *Proceedings of the Conference on Language Modeling (COLM)*.

Jin, W.; Min, T.; Yang, Y.; Wei, D.; Zhou, Y.; Kadhe, S. R.; Baracaldo, N.; and Lee, K. 2026. Entropy-Aware On-Policy Distillation of Language Models. *arXiv preprint arXiv:2603.07079*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1601–1611.

Kang, M.; Jeong, J.; Lee, S.; Cho, J.; and Hwang, S. J. 2025. Distilling LLM Agent into Small Models with Retrieval and Code Tools. *arXiv preprint arXiv:2505.17612*.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769–6781.

Kim, Y.; and Rush, A. M. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1317–1327.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; and Toutanova, K. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*, 7: 452–466.

Lin, W.; and Han, K. 2026. Surgical Post-Training: Proximal On-Policy Distillation for Reasoning with Knowledge Retention. *arXiv preprint arXiv:2603.01683*.

Liu, X.; Jiao, K.; Xiao, C.; Zhao, R.; Ruan, J.; et al. 2026. Teacher-Guided Policy Optimization for On-Policy Reasoning Distillation under Large Policy Divergence. *arXiv preprint arXiv:2605.13230*.

Ma, Y.; Liang, Z.; Chen, B.; Qian, Z.; Dai, H.; Mao, L.; Zhang, X.; Lei, C.; and Ou, W. 2026. SD-Search: On-Policy Hindsight Self-Distillation for Search-Augmented Reasoning. *arXiv preprint arXiv:2605.18299*.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. 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)*, 9802–9822.

Nakano, R.; Hilton, J.; Balwit, A.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. WebGPT: Browser-Assisted Question-Answering with Human Feedback. *arXiv preprint arXiv:2112.09332*.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N. A.; and Lewis, M. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5687–5711.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*, volume 36.

Shan, Y.; Guo, Y.; Cheng, Z.; Liu, Z.; Zhu, X.; et al. 2026. Learning from Own Solutions: Self-Conditioned Credit Assignment for Reinforcement Learning with Verifiable Rewards. *arXiv preprint arXiv:2606.18810*.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2024. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv preprint arXiv:2409.19256*.

Shi, Y.; Li, S.; Wu, C.; Liu, Z.; Fang, J.; Cai, H.; Zhang, A.; and Wang, X. 2025. Search and Refine During Think: Facilitating Knowledge Refinement for Improved Retrieval-Augmented Reasoning. In *Advances in Neural Information Processing Systems*.

Song, H.; Jiang, J.; Min, Y.; Chen, J.; Chen, Z.; Zhao, W. X.; Fang, L.; and Wen, J.-R. 2025. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2503.05592*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association of Computational Linguistics*, 10: 539–554.

Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; and Wei, F. 2024. Improving Text Embeddings with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11897–11916.

Wang, Z.; Zheng, X.; An, K.; Ouyang, C.; Cai, J.; Wang, Y.; and Wu, Y. 2025. StepSearch: Igniting LLMs Search Ability via Step-Wise Proximal Policy Optimization. *arXiv preprint arXiv:2505.15107*.

Xu, H.; Zhu, Q.; Deng, H.; Li, J.; Hou, L.; Wang, Y.; Shang, L.; Xu, R.; and Mi, F. 2025. KDRL: Post-Training Reasoning LLMs via Unified Knowledge Distillation and Reinforcement Learning. *arXiv preprint arXiv:2506.02208*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA:

A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*.

Zhang, Z.; Jiang, S.; Shen, Y.; Zhang, Y.; Ram, D.; Yang, S.; Tu, Z.; Xia, W.; and Soatto, S. 2026. Reinforcement-Aware Knowledge Distillation for LLM Reasoning. *arXiv preprint arXiv:2602.22495*.

Zhao, A.; Xin, H.; Fan, Y.; Tong, J.; Li, W.; and Shen, X. 2026. Decoupling KL and Trajectories: A Unified Perspective for SFT, DAgger, Offline RL, and OPD in LLM Distillation. *arXiv preprint arXiv:2605.16826*.