WHY DOES FEEDBACK-AUGMENTED SELF-DISTILLATION FAIL TO IMPROVE RETRIEVAL-INTERLEAVED SEARCH AGENTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) offers a promising approach for training large language models without relying on a separate teacher model. However, its effectiveness on complex agentic tasks remains largely unexplored. In this work, we instantiate Feedback-Augmented Self-Distillation (FA-SD), a self-distillation algorithm for agentic search that leverages successful demonstrations as privileged information. We observe that models can rely on recurring reasoning-and-search output templates, producing trajectories that appear diverse but are largely agnostic to the input question, making the KL-based self-distillation signal less informative. We term this phenomenon *decoding collapse*, a failure mode that can be missed by aggregate evaluation metrics. To understand why this occurs, we show that although the self-teacher achieves stronger performance, learning remains unstable under inconsistent supervision signals. We further analyze two sources of inconsistency, model inconsistency and prompt inconsistency, and find that prompt-conditioned feedback can degrade the quality of the supervision signal, limiting the effectiveness of self-teacher learning. To mitigate this inconsistency, we introduce an exponential moving average (EMA) teacher to stabilize the self-teacher and provide more consistent supervision signals. Although the EMA teacher requires a warm-up phase during which performance may temporarily regress, it partially improves FA-SD by providing more stable supervision, while leaving a gap to stronger RL and teacher-distillation baselines.

1 INTRODUCTION

Search-augmented language agents interleave reasoning with external retrieval, deciding when to search, how to formulate queries, and how to use retrieved evidence. Reinforcement learning (RL) is a natural fit for this setting because final-answer correctness provides an outcome reward, and recent work shows that RL can improve retrieval-augmented question answering (Jin et al., 2025; Jiang et al., 2025b). However, such rewards are sparse and trajectory-level, providing limited guidance for intermediate search and reasoning decisions. This further motivates the use of denser learning signals for search-agent training, including process rewards, evidence evaluation, and step-level supervision (Li et al., 2026; Shu et al., 2026; Wang & Jiang, 2026).

A complementary source of dense supervision is on-policy distillation (OPD), which trains on student-generated trajectories while using a teacher policy to provide local distributional feedback, reducing the mismatch between fixed teacher traces and student-generated prefixes (Agarwal et al., 2023). Recent on-policy self-distillation (OPSD) methods further suggest that a single LLM can serve as both teacher and student by conditioning the teacher branch on privileged information or self-generated solutions (Zhao et al., 2026; Hübotter et al., 2026). This paradigm turns training-only information into dense token-level supervision without requiring a separate stronger teacher, and is especially natural for search agents because rollout groups may contain both failed and successful attempts for the same question, allowing successful rollouts to serve as privileged demonstrations.

Despite these appealing properties, it remains unclear whether these benefits extend to retrieval-interleaved agents. Search-agent trajectories differ from those considered in standard language-model distillation because model actions determine which evidence is retrieved, while retrieved observations are produced by the environment rather than by the policy (Jin et al., 2025; Li et al., 2026). Consequently, matching token-level distributions may be insufficient to transfer the long-

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.

horizon search decisions required for effective retrieval. The challenge is particularly pronounced for feedback-augmented self-distillation: although successful rollouts may elicit stronger behavior when included in the prompt, training must transfer this contextual advantage to the unconditioned policy. Recent analyses of self-distillation for reasoning similarly suggest that rich teacher conditioning can induce overconfident reasoning behavior that is not automatically beneficial when distilled into an unconditioned model (Kim et al., 2026). This raises the possibility that self-distillation may reduce the discrepancy between teacher and student distributions without producing useful improvements in the resulting search behavior. Moreover, because the self-teacher is both model-dependent and prompt-conditioned, the resulting supervision signal may be inconsistent across training steps and feedback prompts. Whether feedback-augmented self-distillation can provide reliable supervision for retrieval-interleaved search agents therefore remains an open empirical question.

We therefore study whether feedback-augmented self-distillation can provide reliable dense supervision for retrieval-interleaved search agents. We instantiate Feedback-Augmented Self-Distillation (FA-SD), which uses successful rollouts as privileged demonstrations and distills behavior elicited by feedback augmentation into the unconditioned search policy. Across seven open-domain QA benchmarks, FA-SD does not sustain improvement across the evaluated variants despite using successful demonstrations to construct its self-teacher. Our diagnostics show that the failure is not simply due to a weak self-teacher: the feedback-augmented self-teacher can achieve stronger inference-time performance than the unconditioned student. This failure points to *decoding collapse*: models can rely on recurring reasoning-and-search output templates, producing diverse-looking but input-question-agnostic trajectories that make the KL-based self-distillation signal less informative. Such behavior can be missed by standard aggregate performance metrics, which do not directly test whether decoded trajectories are input-specific.

To understand this collapse, we analyze two sources of instability in the FA-SD supervision signal: model inconsistency, caused by the evolving self-teacher, and prompt inconsistency, caused by conditioning the teacher branch on feedback demonstrations. We then evaluate fixed-reference and exponential moving average (EMA) teachers to stabilize self-teacher supervision. Finally, we test whether the feedback-augmentation trick generalizes to external-teacher OPD. Although standard multi-rollout OPD (MOPD) trains more stably with a strong external teacher, Feedback-Augmented MOPD underperforms standard MOPD, suggesting that privileged feedback can introduce prompt inconsistency when paired with an already strong teacher. Our contributions are fourfold. First, we instantiate FA-SD, a feedback-augmented self-distillation algorithm for retrieval-interleaved search agents that uses successful rollouts as privileged demonstrations. Second, we characterize *decoding collapse*, where models can follow recurring reasoning-and-search output templates, producing diverse-looking but input-question-agnostic trajectories that make the KL-based self-distillation signal less informative. Third, we show that fixed-reference and EMA teachers help mitigate model inconsistency by reducing self-teacher drift, thereby stabilizing FA-SD training. Fourth, we show that a strong external teacher provides more stable early supervision in MOPD, yet feedback augmentation does not straightforwardly generalize to external-teacher OPD, suggesting that privileged feedback may introduce prompt inconsistency or unnecessary conditioning.

2 Related Work

Search-agent reinforcement learning. Recent work extends retrieval-augmented generation into search-augmented reasoning agents that interleave reasoning with search or tool-use actions. Search-R1 trains LLMs with RL to issue search queries during step-by-step reasoning (Jin et al., 2025), while related query-optimization and search-agent methods learn retrieval policies or lightweight searchers for downstream question answering (Jiang et al., 2025a;b). Tool-use RL further studies how agents acquire tool-calling behavior with limited or no supervised traces (Acikgoz et al., 2026). These approaches establish RL as an effective framework for improving search and tool-use behavior, but primarily rely on outcome-level or process-level rewards rather than self-distillation from privileged rollout demonstrations.

Dense feedback and self-improving agents. Dense feedback has been proposed to address weak credit assignment in outcome-level rewards. SE-Search combines memory, atomic query training, and dense rewards for search agents (Li et al., 2026); SLEA-RL augments multi-turn RL with step-level experience and credit assignment (Wang & Jiang, 2026); and EvalAct makes evidence evaluation an explicit action to obtain process-aligned rewards (Shu et al., 2026). Related self-improving agents evolve curricula or revise trajectories for iterative improvement (Gao et al., 2025; Xia et al., 2025; Acikgoz et al., 2026; Lin et al., 2025). These methods motivate denser supervision for agent

training, but do not directly examine whether successful rollouts can be reused as privileged demonstrations for token-level self-distillation.

Policy distillation and on-policy self-distillation. Policy distillation provides another way to supply dense supervision. On-policy distillation (OPD) trains on student-generated prefixes while using a teacher policy to provide distributional targets, reducing the mismatch between fixed teacher demonstrations and student-generated trajectories (Agarwal et al., 2023). Recent work identifies practical failure modes in sampled-token OPD, including imbalanced token-level supervision and unreliable teacher guidance on student-generated prefixes, and proposes local support matching to improve training stability (Fu et al., 2026). Instead of distilling from a single response per task, Yu et al. (2026) propose Multi-Rollout On-Policy Distillation (MOPD), which leverages a group of student rollouts for the same task to construct more informative teacher supervision. Complementary on-policy self-distillation (OPSD) methods show that feedback, privileged information, or self-generated solutions can form a self-teacher for reasoning and verifiable tasks (Zhao et al., 2026; Hübotter et al., 2026). However, self-distillation can degrade reasoning when rich teacher conditioning induces overconfident behavior that does not transfer cleanly to the unconditioned model (Kim et al., 2026). These concerns are sharper in retrieval-interleaved agents, where search actions affect retrieved observations and successful rollouts provide privileged demonstrations through prompts rather than fixed labels.

Collapse, teacher stability, and negative results. Collapse is a central concern in bootstrap-style self-supervised learning, where models may learn degenerate representations rather than informative features. BYOL studies this issue in representation learning and uses a slowly moving target network to stabilize bootstrap learning (Grill et al., 2020). In language-model self-distillation and agentic RL, collapse can instead appear in decoding behavior; rich teacher conditioning may degrade reasoning (Kim et al., 2026), while agentic models may produce diverse-looking but input-question-agnostic trajectories from recurring output templates (Wang et al., 2026). Thus, a small distillation loss need not imply useful reasoning or search behavior. Our work also relates to negative results in agent training, where plausible enhancements or richer feedback signals may fail to improve strong baselines. CTIM-Rover adds episodic memory to a software-engineering agent but fails to outperform AutoCodeRover in any configuration (Lindenbauer et al., 2025); search-and-self-feedback results show that model-generated feedback is not a reliable substitute for carefully designed or ground-truth feedback (K et al., 2025). Similar negative findings in NLP show that intermediate-layer distillation or question-rewriting objectives may fail to outperform simpler baselines (Suzuki et al., 2025; Ishii et al., 2022). These findings motivate diagnosing why richer feedback signals fail, rather than reporting performance regressions alone.

3 METHOD

We formulate dense distillation over retrieval-interleaved search-agent rollouts, where the policy interleaves reasoning, search actions, retrieved observations, and final-answer generation. Under this rollout formulation, we study two sources of token-level supervision: teacher-based on-policy distillation (OPD) from an external teacher, and feedback-augmented self-distillation (FA-SD) from rollout-derived feedback.

3.1 SEARCH-AGENT ROLLOUT SETUP

Given a question x from a dataset $\mathcal{S}$ and a search engine $\mathcal{R}$, the policy π_θ with parameters θ generates a multi-step trajectory:

$$\tau = (r_1, s_1, o_1, r_2, s_2, o_2, \ldots, r_K, s_K, o_K, a), \tag{1}$$

where $k = 1, \ldots, K$ indexes retrieval-interleaved steps, K is the total number of interaction steps, r_k and s_k denote the model-generated reasoning and search-query token sequences at step k, $o_k = \mathcal{R}(s_k)$ denotes the observation retrieved when the model invokes search, and a is the final answer. A task reward $R(x, \tau)$ is computed from the final answer, e.g., exact match or LLM judge score.

3.2 DISTILLATION WITH A SEARCH ENGINE

We consider two distillation formulations with a search engine: teacher-based OPD, where an external teacher supervises student-generated trajectories, and FA-SD, where the policy is conditioned on the augmented feedback to form a self-teacher.

3.2.1 TEACHER-BASED ON-POLICY DISTILLATION WITH A SEARCH ENGINE

For teacher-based on-policy distillation, student-generated retrieval-interleaved trajectories are supervised by an external teacher policy at each token position along the sampled trajectory:

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{S},\, \hat{y}\sim p_S(\cdot|x;\mathcal{R})}\left[\frac{1}{|\hat{y}|}\sum_{t=1}^{|\hat{y}|}\mathcal{D}\right.$$
$$\left.\left(p_T(\cdot \mid x, \hat{y}_{<t}; \mathcal{R}) \,\|\, p_S(\cdot \mid x, \hat{y}_{<t}; \mathcal{R})\right)\right]. \tag{2}$$

where $\mathcal{S}$ is the dataset, $\hat{y}$ denotes the token sequence of a sampled student trajectory, $\hat{y}_{<t}$ denotes its prefix before position t, and $|\hat{y}|$ is the sequence length. The policies p_S and p_T denote the student and external teacher models, respectively, with p_S corresponding to the trainable policy π_θ in the teacher-distillation objective. $\mathcal{D}$ denotes a divergence measure, such as the Kullback–Leibler (KL) divergence or the reverse KL divergence (Agarwal et al., 2023). Unlike prior OPD approaches that rely on the policy itself $p(\hat{y}|x)$, our formulation explicitly incorporates retrieval-interleaved reasoning through $p(\hat{y}|x;\mathcal{R})$, which can be viewed as coupling the policy with the search engine $\mathcal{R}$, conceptually denoted as $p(\hat{y}|x)\otimes\mathcal{R}$, to model interleaved retrieval and reasoning.

3.2.2 Feedback-Augmented Self-Distillation with a Search Engine

Alternatively, motivated by the self-teacher ideas in recent on-policy self-distillation methods (Hübotter et al., 2026; Zhao et al., 2026), we instantiate a feedback-augmented self-distillation objective for retrieval-interleaved search-agent rollouts. We refer to this formulation as Feedback-Augmented Self-Distillation (FA-SD), since it preserves the OPD-style dense token-level supervision objective while replacing the external teacher with the policy itself conditioned on rollout-augmented feedback.

Here, feedback refers to additional information that may help the policy solve the same question. In our search-agent setting, the feedback f is a successful solution sampled from another attempt on x within the same rollout group. The self-teacher is then defined as $p_{\text{ST}}(\cdot \mid x, f; \mathcal{R}) := \pi_\theta(\cdot \mid x, f; \mathcal{R})$, representing the current policy prompted with the question x and feedback f. The loss function is then derived as

$$\mathcal{L}_{\text{FA-SD}}(\theta) = \mathbb{E}_{x\sim\mathcal{S},\, \hat{y}\sim\pi_\theta(\cdot|x;\mathcal{R})}\left[\frac{1}{|\hat{y}|}\sum_{t=1}^{|\hat{y}|}\right.$$
$$\left.\mathcal{D}\Big(\pi_\theta^t \,\|\, \text{stopgrad}(\pi_\theta^{f,t})\Big)\right], \tag{3}$$

with

$$\pi_\theta^{f,t} = \pi_\theta(\cdot \mid x, \hat{y}_{<t}, f; \mathcal{R}), \quad \pi_\theta^t = \pi_\theta(\cdot \mid x, \hat{y}_{<t}; \mathcal{R}),$$

where π_θ^t is the unconditioned student token distribution at position t, and $\pi_\theta^{f,t}$ is the feedback-augmented self-teacher token distribution at the same prefix. The stopgrad operator blocks gradients through the self-teacher branch, and thus this branch serves as a per-step target for the student and preserves the supervision signal provided by f. Specifically, for each question x, we first sample a rollout group of G trajectories. If there exists at least one correct trajectory, we reprompt the samples by augmenting the prompt with one randomly selected correct trajectory (excluding itself), using the feedback-augmentation template in Appendix A. The loss above corresponds to the (unclipped) FA-SD implementation. For the reverse-KL instantiation of $\mathcal{D}$ in Eq. 3, following Hübotter et al. (2026), the gradient is derived as

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{\text{FA-SD}} &= \mathbb{E}_{x,\hat{y}}\left[\frac{1}{|\hat{y}|}\sum_t G_t^{\text{FA-SD}}\right], \\
G_t^{\text{FA-SD}} &= \mathbb{E}_{z\sim\pi_\theta^t}\left[\log \rho_t^{\text{FA-SD}}(z)\, \nabla_\theta \log \pi_\theta^t(z)\right], \\
\log \rho_t^{\text{FA-SD}}(z) &= \log \pi_\theta^t(z) - \log \pi_\theta^{f,t}(z),
\end{aligned} \tag{4}$$

where $G_t^{\text{FA-SD}}$ denotes the per-token gradient contribution at position t, z denotes a candidate assistant token sampled from the unconditioned student distribution π_θ^t, and $\rho_t^{\text{FA-SD}}(z) = \pi_\theta^t(z)/\pi_\theta^{f,t}(z)$ is the student-to-self-teacher probability ratio. The distributions π_θ^t and $\pi_\theta^{f,t}$ are defined above, and gradients do not flow through the feedback-augmented self-teacher branch.

Under gradient descent, if the student assigns higher probability than the self-teacher to token z, then $\log \rho_t^{\text{FA-SD}}(z) > 0$ and the update decreases the student's probability on that token. If the student assigns lower probability than the self-teacher, the log ratio is negative and the update increases the student's probability. Because retrieved observations are environment outputs rather than model actions, we mask retrieved passages and compute the FA-SD loss only on assistant-generated tokens.

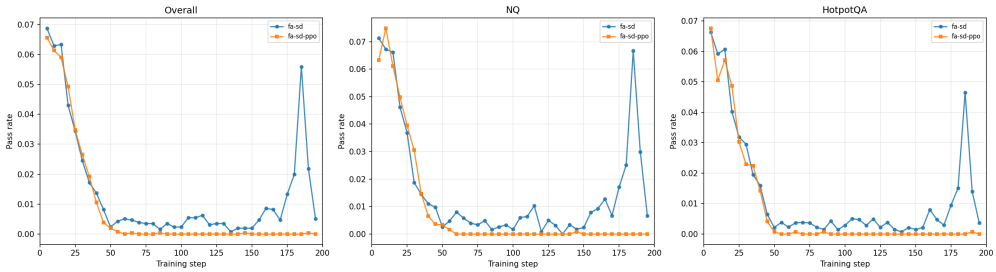

Figure 1: Training pass-rate dynamics for unclipped and PPO-style clipped FA-SD variants over all sampled rollouts (Overall) and by dataset (NQ, HotpotQA). Both variants fail to sustain improvement: pass rates remain close to zero for much of training, with only a transient rebound around step 180 before returning to near-zero levels.

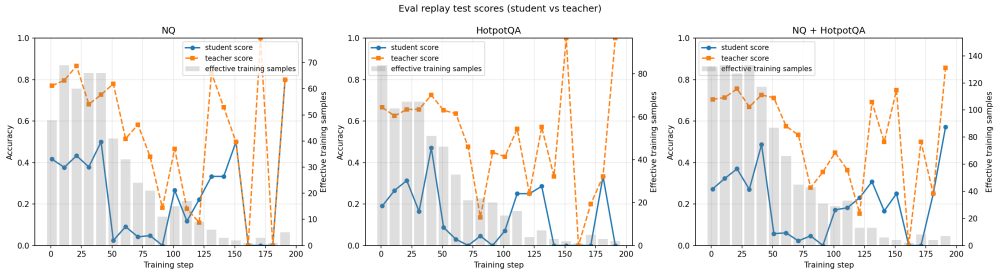

Figure 2: Evolution of inference-time accuracy of the unconditioned student and feedback-augmented self-teacher during FA-SD training. Gray bars show effective training-sample counts. The self-teacher is often stronger than the student, but this advantage is not reliably internalized.

We also evaluate a PPO-style clipped FA-SD variant to test whether more conservative policy updates improve stability, as well as a joint FA-SD+GRPO variant that combines token-level self-distillation with outcome-level RL rewards using the same rollout data; further implementation details are provided in Appendix B.

3.3 Regularization on Self-Teacher

In naive FA-SD, the feedback-augmented self-teacher and the unconditioned student are parameterized by the same continuously updated policy. As a result, the teacher target can drift during training, producing inconsistent supervision and unstable optimization, especially in long-horizon agentic tasks such as search. To help reduce this model inconsistency, we evaluate two self-teacher regularization strategies: a fixed-reference teacher and an exponential moving average (EMA) teacher, which stabilize the self-teacher by preventing or slowing teacher drift.

The fixed-reference strategy uses the student policy at the start of FA-SD training, prompted with the same rollout-derived feedback throughout training. This prevents the teacher target from changing with every student update. The EMA strategy instead uses an exponential moving average teacher, whose parameters are updated as

$$\theta_t^{\mathrm{EMA}} = \alpha\theta_{t-1}^{\mathrm{EMA}} + (1-\alpha)\theta_t, \tag{5}$$

where θ_t denotes the student parameters at training step t, and $\alpha \in [0, 1)$ is the EMA momentum coefficient. The EMA teacher is initialized as $\theta_0^{\mathrm{EMA}} = \theta_0$. A larger α causes the teacher to adapt more slowly to the student policy, resulting in a more stable teacher. Unless otherwise specified, we set $\alpha = 0.95$ in all experiments.

3.4 Feedback-Augmented External Teacher

In our experiments, we find that self-distillation requires an initial warm-up phase, during which the model may temporarily regress before the self-teacher becomes sufficiently stable to provide consistent distillation signals. To further understand the effect of the feedback-augmentation trick, we extend it to the multi-rollout on-policy distillation (MOPD) framework. Specifically, we augment the external teacher's prompt with a successful rollout, following the same feedback-augmentation strategy introduced in Section 3.2.2 for FA-SD. Unlike FA-SD, however, the augmented prompt is provided to the external teacher, while the student remains unconditioned. We refer to this variant as Feedback-Augmented MOPD (FA-MOPD).

4 Experiments

We conduct experiments to study whether Feedback-Augmented Self-Distillation (FA-SD) can provide reliable dense supervision for retrieval-interleaved search agents. All trainable student check-

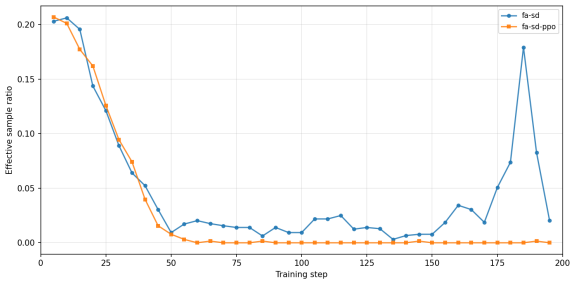

Figure 3: Effective distillation-signal ratio for different FA-SD variants during training. The effective sample ratio closely tracks the training pass rate, suggesting a strong positive correlation between the availability of usable distillation signals and model performance.

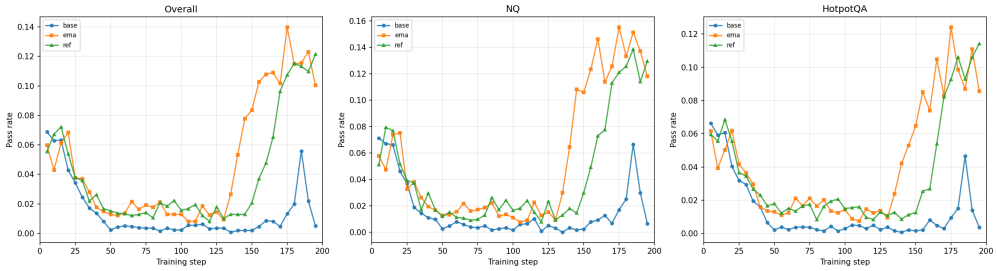

Figure 4: Training pass-rate dynamics of FA-SD with different self-teacher regularization strategies. The baseline FA-SD suffers from unstable optimization and quickly regresses after an initial improvement. In contrast, regularizing the self-teacher with either an EMA teacher or a reference teacher stabilizes training, allowing the model to recover from temporary regressions and achieve higher final performance among FA-SD variants across Overall, NQ, and HotpotQA.

points are based on Qwen2.5-3B-Base. We focus on FA-SD variants that construct a self-teacher from successful rollouts, and include GRPO-trained and teacher-based OPD models as comprehensive comparison points.

Specifically, we conduct comprehensive experiments to answer the following research questions:

(1) How do different variants of FA-SD perform on agentic tasks? (§4.2)

(2) Does the self-teacher in FA-SD outperform the student on agentic tasks? (§4.3)

(3) Why does FA-SD degrade? Does teacher regularization help? (§4.4; §4.6)

(4) Can the feedback-augmentation trick be generalized to the OPD framework? (§4.7)

4.1 Experimental Setup

Datasets and metrics. We follow the open-domain QA evaluation setting used in prior retrieval-augmented search-agent work (Jin et al., 2025; Shu et al., 2026; Li et al., 2026). The training data are formed by merging the training sets of Natural Questions (NQ) and HotpotQA. We evaluate on seven benchmarks: NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle. Following prior work, we report Exact Match (EM) on each benchmark and the average score across all seven tasks. For training dynamics, pass rate denotes the average final-answer task reward over sampled trajectories.

Retrieval corpus and models. We use the 2018 Wikipedia dump as the external knowledge corpus and E5 (Wang et al., 2022) as the retriever, returning the top 3 passages for each search query (Jin et al., 2025; Li et al., 2026). All trainable student policies are initialized from Qwen2.5-3B-Base or from a GRPO-trained checkpoint derived from this base model. For teacher-based OPD, Qwen2.5-7B-Instruct is used as the external teacher to construct dense supervision targets and is also evaluated directly as a teacher reference. For FA-SD, the dense supervision signal is constructed by augmenting the current policy with rich feedback derived from successful rollouts in the same rollout group, rather than from an external teacher.

Training variants. Our main experiments evaluate FA-SD variants, including unclipped FA-SD, PPO-style clipped FA-SD, FA-SD with joint GRPO training, and FA-SD with teacher regularization. We additionally include GRPO-trained and teacher-based OPD models as comprehensive comparison points in Appendix C.5. Additional training details are provided in Appendix B.

4.2 FA-SD Variants Do Not Sustain Improvement

We first evaluate the main FA-SD setting, where successful rollouts are used as privileged demonstrations to construct a feedback-augmented self-teacher. The goal is to test whether this self-

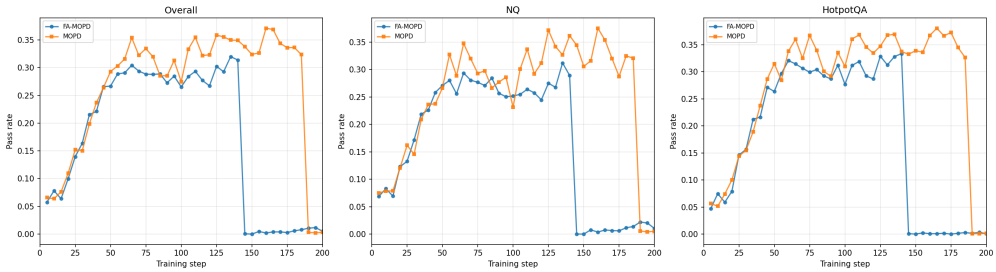

Figure 5: Training pass-rate dynamics for MOPD and Feedback-Augmented MOPD. Standard MOPD improves more steadily during early training without the temporary regression observed in regularized FA-SD, suggesting more stable early external-teacher supervision; FA-MOPD underperforms standard MOPD, indicating that privileged feedback does not straightforwardly generalize to external-teacher OPD.

teacher can provide dense token-level supervision that improves the unconditioned search policy. We consider the FA-SD variants described in Appendix B, including unclipped FA-SD and PPO-style clipped FA-SD. As shown in Figure 1, both variants fail to sustain improvement: their pass rates remain close to zero for much of training, with only a transient rebound around step 180 before returning to near-zero levels. We therefore focus the following sections on diagnosing why the feedback-augmented self-teacher signal fails to provide stable and informative supervision.

4.3 Feedback-Augmented Self-Teacher Advantage Is Not Reliably Internalized

We next ask whether FA-SD fails because the feedback-augmented self-teacher is too weak. To test this, we directly compare the unconditioned student branch and the feedback-augmented self-teacher branch at inference time, without applying parameter updates.

Figure 2 shows that the feedback-augmented self-teacher branch is often stronger than the unconditioned student. This rules out the simplest explanation that FA-SD fails only because the self-teacher lacks useful capability. However, the student does not reliably catch up to the self-teacher branch over training. This suggests an internalization failure: rollout-derived feedback can improve behavior when provided in the prompt, but the resulting token-level distillation signal is not reliably absorbed into the unconditioned search agent.

4.4 Decoding Collapse and Supervision Inconsistency

We next diagnose why FA-SD degrades despite a stronger feedback-augmented self-teacher. We refer to this failure mode as *decoding collapse*: models rely on recurring reasoning-and-search output templates, producing diverse-looking but input-question-agnostic trajectories that make the KL-based self-distillation signal less informative. In our experiments, this collapse is reflected by a failure to sustain pass-rate improvement and an increasingly sparse effective distillation signal.

As shown in Figure 1, both unclipped FA-SD and PPO-style clipped FA-SD exhibit near-zero pass rates for much of training. This degraded training dynamic motivates a closer examination of whether the feedback-augmented self-distillation objective continues to provide usable supervision. Figure 3 provides this complementary diagnostic. We measure the effective sample ratio as the fraction of sampled trajectories that provide a valid feedback-augmented distillation signal. For both FA-SD variants, this ratio decreases during training, indicating that the usable self-distillation signal becomes increasingly sparse. Together with the inference-time diagnostic in Figure 2, this suggests that the bottleneck is not simply whether the feedback-augmented self-teacher can exhibit stronger behavior, but whether that behavior can be converted into sufficiently dense and stable supervision for the unconditioned student. Additional per-variant training dynamics are provided in Appendix C.2. Moreover, we conducted an experiment combining the FA-SD and GRPO objectives. We observed that both the training pass rate and validation performance degraded, suggesting that FA-SD introduces inconsistent supervision signals that interfere with and ultimately hinder the standard GRPO training process.

4.5 Behavioral Diagnostics of Decoding Collapse

The preceding analysis motivates a direct examination of the behavioral signatures of decoding collapse. We therefore evaluate three diagnostics of trajectory input specificity, rather than relying only on aggregate pass-rate trends. Specifically, we diagnose behavioral collapse along three complementary axes, computed offline from per-step training rollouts. For each step, we remove the retrieved blocks enclosed by <information>...</information> from the assistant text, and then extract search

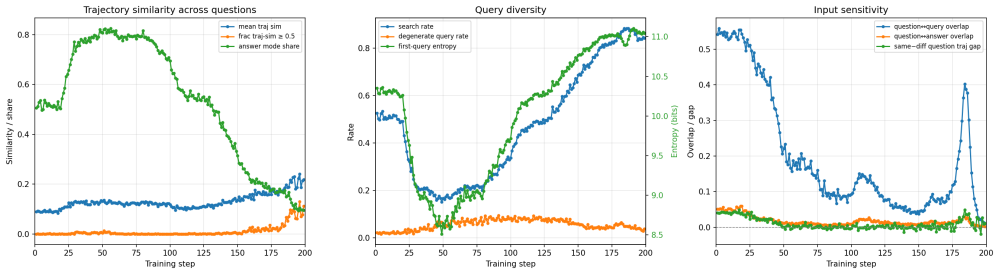

Figure 6: Behavioral diagnostics of decoding collapse for FA-SD across training steps.

queries enclosed by <search>...</search> and final answers enclosed by <answer>...</answer>. We describe the three metric categories below.

Trajectory similarity. We measure the similarity between assistant trajectories using four-character n-gram Jaccard. At each step, we report three metrics: (i) mean trajectory similarity over 256 subsampled cross-example pairs; (ii) the fraction of pairs with similarity ≥ 0.5 (near-duplicate trajectories); and (iii) answer mode share, the fraction of rollouts whose answer, extracted from <answer>...</answer>, equals the most frequent answer at that step. Collapse appears as rising mean similarity, a growing high-similarity mass, and a dominant answer mode.

Query diversity. We measure search behavior using three metrics: (i) search rate, the fraction of rollouts that issue at least one search; (ii) first-query entropy, the Shannon entropy (bits) of the empirical distribution over first queries at that step; and (iii) degenerate-query rate, the fraction of queries that are too short or too long, or match known recovery boilerplate (e.g., "my previous action is invalid"). Details of the quantitative metrics are provided in Appendix C.4. Collapse appears as falling search rate and first-query entropy, and/or rising degenerate-query rate.

Input sensitivity. We analyze how outputs depend on the input question. Lexically, we compute stopword-filtered token recall of the question in (i) the search queries (question $\leftrightarrow$ query overlap) and (ii) the final answer (question $\leftrightarrow$ answer overlap). Behaviorally, we compare mean trajectory similarity among same-question rollout pairs versus different-question pairs, and define the conditioned gap as same-question similarity minus different-question similarity, measured by four-character n-gram Jaccard. Details are provided in Appendix C.4. A healthy policy should maintain a positive gap, indicating that matched inputs produce more similar trajectories than mismatched inputs.

We plot these metrics across the FA-SD training steps in Figure 6. The trajectory-similarity diagnostics show that collapse first appears as answer-mode locking and later as cross-question trajectory homogenization. The query-diversity metrics show that search behavior remains present, but collapse occurs in the middle stage with reduced search rate and lower query diversity. In the late stage, the search rate and entropy rebound, but this does not restore QA behavior. Finally, the input-sensitivity plot shows that question-query overlap closely tracks pass-rate trends, while the model becomes largely question-insensitive after 50 steps, suggesting that the policy has entered a collapsed state.

4.6 Regularization on Teacher to Stabilize Optimization

The previous diagnostics show that FA-SD fails to sustain pass-rate improvement and that its effective distillation signal becomes increasingly sparse, despite the presence of a stronger feedback-augmented self-teacher. We therefore ask whether stabilizing the self-teacher can make the supervision signal more consistent. Instead of allowing the self-teacher to share the same continuously updated weights as the student, we regularize the self-teacher using either a fixed reference model or an exponential moving average (EMA) of the student model. These strategies either stop or slow drift in the teacher parameters, thereby providing more consistent supervision throughout training. Figure 4 shows that both strategies stabilize FA-SD training. A fixed reference teacher improves the consistency of the distillation signal, allowing the model to recover and continue improving after a temporary performance regression. The EMA teacher achieves faster recovery and converges to higher final performance across Overall, NQ, and HotpotQA, suggesting that gradual teacher adaptation provides a better balance between stability and adaptability. The EMA-regularized FA-SD checkpoint obtains an average score of 0.206 across the seven evaluation benchmarks, improving over unregularized FA-SD while remaining below stronger GRPO and OPD baselines; full EMA per-task results and additional comparisons are provided in Appendices C.3 and C.5. Compared with the initial Qwen3B-Base average score of 0.114 reported in Appendix C.3, FA-SD with EMA regularization improves the overall final evaluation score.

On the other hand, the behavior in Figure 4 suggests that the self-teacher requires an initial warm-up phase, during which the model may experience temporary performance regression before the teacher provides sufficiently consistent distillation signals. We hypothesize that this phenomenon arises because the self-teacher is initially too weak to produce useful learning signals with good generalization, making those signals difficult for the student to learn from effectively.

4.7 Generalizing the Feedback-Augmentation Trick to OPD

We begin by testing the above hypothesis. Specifically, we repeat the same experiment by replacing the self-teacher with an external teacher, following the standard multi-rollout on-policy distillation (MOPD) framework, where multiple rollouts collected for the same task are leveraged to perform policy distillation. To ensure a fair comparison, we use the same rollout group size as in FA-SD and keep all other training settings unchanged. Figure 5 shows that MOPD improves more steadily during early training without experiencing the temporary performance regression observed in regularized FA-SD. This result supports our hypothesis that a strong external teacher provides more generalizable and consistent early supervision, enabling the student to benefit from the distillation objective in the early stages of training.

Moreover, we incorporate the feedback augmentation strategy into the MOPD framework by augmenting the teacher prompt with a successful rollout, following the feedback-augmentation template in Appendix A. Unlike FA-SD, however, the augmented prompt is provided to an external teacher rather than the self-teacher. We refer to this variant as Feedback-Augmented MOPD (FA-MOPD) and report its training pass rate in Figure 5. The results show that FA-MOPD underperforms the standard MOPD baseline, indicating that feedback augmentation alone is not sufficient to improve external-teacher distillation in this setting. These results suggest that, when the external teacher already provides strong supervision, incorporating privileged feedback may introduce prompt inconsistency or unnecessary conditioning rather than providing useful additional supervision.

4.8 Discussion

Our experiments show that feedback-augmented self-distillation does not automatically provide reliable dense supervision for retrieval-interleaved search agents. Although successful rollouts can serve as privileged demonstrations for a feedback-augmented self-teacher, FA-SD does not sustain improvement across variants. Together with the GRPO and OPD comparisons in Appendix C.5, these results suggest that dense token-level distributional supervision does not, by itself, guarantee reliable search-policy improvement.

The diagnostics show that FA-SD fails not simply because the self-teacher is weak: the feedback-augmented self-teacher can achieve stronger inference-time performance, but its signal is not reliably internalized, consistent with prior findings that rich teacher conditioning may not transfer cleanly (Kim et al., 2026). Instead, FA-SD exhibits *decoding collapse*: diverse-looking trajectories can follow recurring reasoning-and-search templates and remain input-question-agnostic, making the KL-based signal less informative.

We interpret this collapse as associated with inconsistent self-teacher supervision across model updates and feedback prompts. Successful rollout feedback can make the self-teacher behave as if useful evidence has already been identified, while the unconditioned student must still discover such evidence through search. Teacher regularization with fixed-reference or EMA teachers helps stabilize FA-SD, suggesting that self-teacher instability is a key factor. At the same time, FA-MOPD underperforms standard MOPD, suggesting that privileged feedback may also introduce prompt inconsistency or unnecessary conditioning when paired with an already strong teacher.

5 Conclusion

We studied whether Feedback-Augmented Self-Distillation (FA-SD) can provide reliable dense supervision for retrieval-interleaved search agents. Although successful rollouts offer privileged demonstrations, FA-SD does not sustain improvement across variants. The failure is not simply due to a weak self-teacher: the feedback-augmented self-teacher can achieve stronger inference-time performance, but its supervision is not reliably internalized. We characterize *decoding collapse* as a key failure pattern and interpret it as associated with inconsistent supervision across model updates and feedback prompts. Fixed-reference and EMA teachers stabilize training but do not fully resolve this difficulty. These findings suggest that richer self-generated supervision alone does not guarantee reliable improvement; future work should determine what to distill from successful rollouts, when feedback-augmented teachers should be trusted, and how to stabilize self-distillation in retrieval-interleaved agents.

AI USE STATEMENT

We did not use generative AI tools for tasks with required disclosure under the ICLR 2027 AI Policy for Authors. We used generative AI tools for the recommended-disclosure task of aiding and polishing writing, including improving clarity, grammar, wording, and organization of the manuscript. All AI-assisted edits were reviewed and revised by the authors. We take responsibility for the final content of this work, including any text revised with the aid of generative AI.

REFERENCES

Emre Can Acikgoz, Cheng Qian, Jonas Hübotter, Heng Ji, Dilek Hakkani-Tür, and Gokhan Tur. Tool-R0: Self-evolving LLM agents for tool-learning from zero data. *arXiv preprint arXiv:2602.21320*, 2026. doi: 10.48550/arXiv.2602.21320. URL https://arxiv.org/abs/2602.21320.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. *arXiv preprint arXiv:2306.13649*, 2023. doi: 10.48550/arXiv.2306.13649. URL https://arxiv.org/abs/2306.13649.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026. doi: 10.48550/arXiv.2603.25562. URL https://arxiv.org/abs/2603.25562.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, Qihan Ren, Cheng Qian, Zhenhailong Wang, Minda Hu, Huazheng Wang, Qingyun Wu, Heng Ji, and Mengdi Wang. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025. doi: 10.48550/arXiv.2507.21046. URL https://arxiv.org/abs/2507.21046.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. doi: 10.48550/arXiv.2601.20802. URL https://arxiv.org/abs/2601.20802.

Etsuko Ishii, Yan Xu, Samuel Cahyawijaya, and Bryan Wilie. Can question rewriting help conversational question answering? In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pp. 94–99, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.insights-1.13. URL https://aclanthology.org/2022.insights-1.13/.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. DeepRetrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*, 2025a. doi: 10.48550/arXiv.2503.00223. URL https://arxiv.org/abs/2503.00223.

Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. s3: You don't need that much data to train a search agent via RL. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21599–21617, Suzhou, China, November 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1095. URL https://aclanthology.org/2025.emnlp-main.1095/.

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. doi: 10.48550/arXiv.2503.09516. URL https://arxiv.org/abs/2503.09516.

Karthikeyan K, Michelle Yuan, Elman Mansimov, Katerina Margatina, Anurag Pratik, Daniele Bonadiman, Monica Sunkara, Yi Zhang, and Yassine Benajiba. A study on leveraging search and self-feedback for agent reasoning. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 259–271, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.realm-1.18. URL https://aclanthology.org/2025.realm-1.18/.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms? *arXiv preprint arXiv:2603.24472*, 2026. doi: 10.48550/arXiv.2603.24472. URL https://arxiv.org/abs/2603.24472.

Jian Li, Yizhang Jin, Dongqi Liu, Hang Ding, Jiafu Wu, Dongsheng Chen, Yunhang Shen, Yulei Qin, Ying Tai, Chengjie Wang, Xiaotong Yuan, and Yabiao Wang. SE-search: Self-evolving search agent via memory and dense reward. *arXiv preprint arXiv:2603.03293*, 2026.

Jiaye Lin, Yifu Guo, Yuzhen Han, Sen Hu, Ziyi Ni, Licheng Wang, Mingguang Chen, Hongzhang Liu, Ronghao Chen, Yangfan He, Daxin Jiang, Binxing Jiao, Chen Hu, and Huacan Wang. SE-Agent: Self-evolution trajectory optimization in multi-step reasoning with LLM-based agents. *arXiv preprint arXiv:2508.02085*, 2025. doi: 10.48550/arXiv.2508.02085. URL https://arxiv.org/abs/2508.02085.

Tobias Lindenbauer, Georg Groh, and Hinrich Schuetze. From knowledge to noise: CTIM-rover and the pitfalls of episodic memory in software engineering agents. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 411–427, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.realm-1.30. URL https://aclanthology.org/2025.realm-1.30/.

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

Takeshi Suzuki, Hiroaki Yamada, and Takenobu Tokunaga. Aligning sizes of intermediate layers by LoRA adapter for knowledge distillation. In *The Sixth Workshop on Insights from Negative Results in NLP*, pp. 100–105, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.insights-1.10. URL https://aclanthology.org/2025.insights-1.10/.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Prince Zizhuang Wang and Shuli Jiang. SLEA-RL: Step-level experience augmented reinforcement learning for multi-turn agentic training. *arXiv preprint arXiv:2603.18079*, 2026. doi: 10.48550/arXiv.2603.18079. URL https://arxiv.org/abs/2603.18079.

Zihan Wang, Chi Gui, Xing Jin, Qineng Wang, Licheng Liu, Kangrui Wang, Shiqi Chen, Linjie Li, Zhengyuan Yang, Pingyue Zhang, Yiping Lu, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. Ragen-2: Reasoning collapse in agentic rl, 2026. URL https://arxiv.org/abs/2604.06268.

Peng Xia, Kaide Zeng, Jiaqi Liu, Can Qin, Fang Wu, Yiyang Zhou, Caiming Xiong, and Huaxiu Yao. Agent0: Unleashing self-evolving agents from zero data via tool-integrated reasoning. *arXiv preprint arXiv:2511.16043*, 2025. doi: 10.48550/arXiv.2511.16043. URL https://arxiv.org/abs/2511.16043.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures. *arXiv preprint arXiv:2605.12652*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. doi: 10.48550/arXiv.2601.18734. URL https://arxiv.org/abs/2601.18734.

APPENDIX

A FEEDBACK-AUGMENTATION TEMPLATE

Table 1 shows the template used to construct the feedback-augmented self-teacher prompt. The placeholder successful_previous_rollout denotes a successful rollout sampled from another attempt on the same question.

User:	prompt Correct solution: successful_previous_rollout Correctly solve the original question.
Assistant:	original_response

Table 1: Template for feedback augmentation.

B TRAINING VARIANTS AND IMPLEMENTATION DETAILS

All trainable student checkpoints are based on Qwen2.5-3B-Base. We evaluate three classes of training variants. First, we train a GRPO baseline in the retrieval-augmented search-agent environment; this checkpoint serves as the main reinforcement-learning baseline. Second, we evaluate teacher-based OPD using Qwen2.5-7B-Instruct as the external teacher. Teacher-based OPD is applied both from the Qwen2.5-3B-Base initialization and from the GRPO-trained checkpoint, allowing us to test whether dense teacher supervision is useful before or after reinforcement learning. Third, we evaluate FA-SD, which replaces the external teacher with a feedback-augmented self-teacher constructed from rollout-derived feedback.

For FA-SD, we consider three variants. The unclipped variant directly optimizes the dense self-distillation objective in Equation 3. The clipped variant optimizes the FA-SD loss through the following PPO-style clipped objective:

$$\mathcal{L}_{\text{FA-SD-PPO}} = \mathbb{E}_{x,\hat{y}}\left[\frac{1}{|\hat{y}|}\sum_t \min\left(r_t(\theta)A_t^{\text{FA-SD}}, \bar{r}_t(\theta)A_t^{\text{FA-SD}}\right)\right], \tag{6}$$

where

$$\begin{aligned} \bar{r}_t(\theta) &= \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\,, \\ A_t^{\text{FA-SD}} &= \log\frac{\pi_\theta^{f,t}(a)}{\pi_\theta^t(a)}, \qquad r_t(\theta) = \frac{\pi_\theta^t(a)}{\pi_{\text{old}}^t(a)}. \end{aligned} \tag{7}$$

The joint variant combines FA-SD with GRPO by reusing the same rollout data to compute both the reinforcement-learning objective and the dense self-distillation objective. Specifically, we always optimize the GRPO loss and adaptively augment it with the FA-SD loss whenever a valid feedback-augmented self-distillation signal is available. In this way, the model can exploit both sparse reward feedback and dense token-level supervision within a unified training framework.

C EXPERIMENTAL DETAILS

C.1 TRAINING HYPERPARAMETERS

The training hyperparameters for GRPO and the FA-SD variants are listed in Tables 2, 3, and 4.

C.2 ADDITIONAL FA-SD TRAINING DYNAMICS

Figures 7 and 8 provide detailed training dynamics for the unclipped FA-SD and PPO-style clipped FA-SD variants, respectively. These figures complement the main-text comparison by showing the reward/pass-rate trend, effective sample ratio, and average number of turns for each variant.

Category	Parameter	Value
Data	Max. prompt length	4096
	Max. response length	500
	Max. start length	2048
	Max. obs, length	500
Search Setup	Max. turns	4
	Retriever topk	3
	Retriever Model	E5
	Retriever Data	Wiki 18
Batching	Question batch size	512
	Mini batch size	256
	Micro batch size	64
	Number of rollouts	5
Rollout	Inference engine	vllm
	Temperature	1.0
Training	Optimizer	AdamW
	Learning rate	1e-6
	Warmup step ratio	0.285
	Max steps	200 (self teacher) or 1000 (external teacher)

Table 2: Common hyperparameters shared by GRPO and FA-SD.

Category	Parameter	Value
Loss	KL loss coef	0.001
	Clip value	0.2

Table 3: GRPO-specific (including PPO-style FA-SD) hyperparameters.

C.3 Final Evaluation of EMA-Regularized FA-SD

Table 5 reports the final evaluation of the EMA-regularized FA-SD checkpoint. We include this result to complement the training-dynamics analysis in Figure 4.

C.4 Collapse Metrics

For degenerate-query rate, A query is labeled degenerate if any of the following hold:

- Length extremes: fewer than 2 characters, or more than 200 characters.
- Boilerplate / recovery text: case-insensitive match to known failure loops, e.g. ”my previous action is invalid”, ”put the query between”, ”let me try again”.

For the same–different question trajectory gap, we compute the similarity metric for same-question trajectories using all 5,120 same-question pairs, and for different-question trajectories using 2,000 randomly sub-sampled different-question pairs.

Category	Parameter	Value
Loss	Distillation divergence	Reverse KL
	KL estimator	Single-sample KL estimator

Table 4: FA-SD specific hyperparameters.

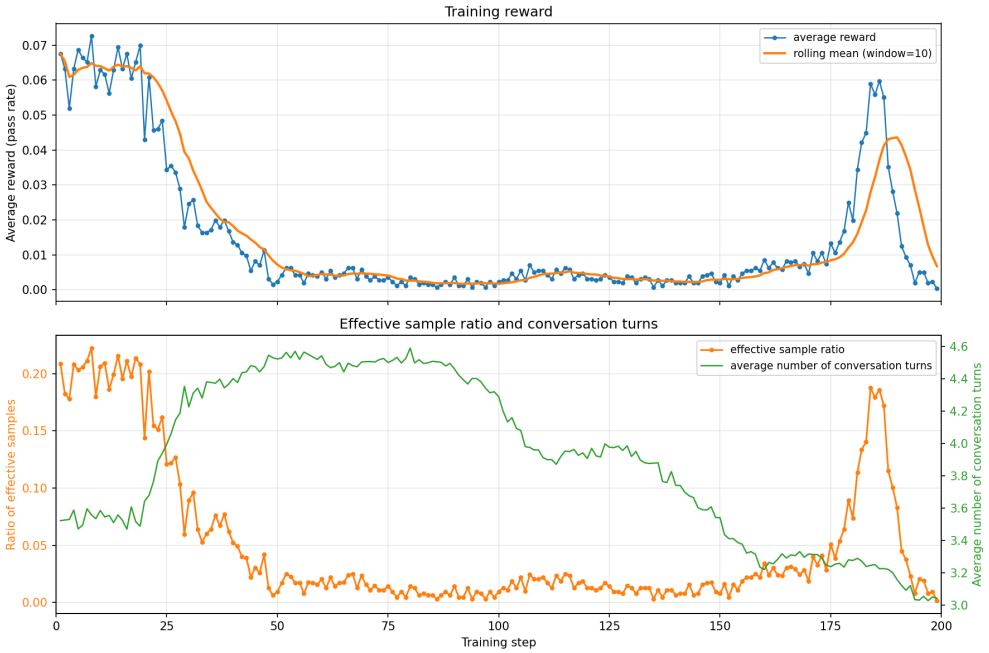

Figure 7: Detailed training dynamics for the unclipped FA-SD variant. The top panel shows the training reward/pass-rate trend, and the bottom panel shows the effective sample ratio together with the average number of turns during training.

C.5 Additional Comparisons with GRPO and Teacher-Based OPD

For completeness, we include additional comparisons with a GRPO-trained search agent and teacher-based OPD variants using Qwen2.5-7B-Instruct as the external teacher. Table 6 reports the full results.

The GRPO-trained Qwen3B checkpoint, initialized from Qwen2.5-3B-Base, obtains an average score of 0.348 across the seven evaluation tasks. This provides an RL-trained reference point for contextualizing the FA-SD results, without serving as the main focus of our analysis.

For teacher-based OPD, we consider the two settings used in our comparison: applying OPD directly to the Qwen3B base model and continuing OPD from the GRPO-trained checkpoint. OPD from the base initialization obtains an average score of 0.274, remaining below the GRPO-trained student. OPD after GRPO improves Bamboogle but reduces performance on the other six benchmarks, decreasing the average score from 0.348 to 0.287. These results suggest that external-teacher token-level supervision does not consistently translate into a stronger retrieval-augmented student in this setting.

D Example Student and Self-Teacher Prompts

We provide an example from HotpotQA to illustrate the construction of the feedback-augmented self-teacher input. The student receives only the original search-agent prompt and question. The

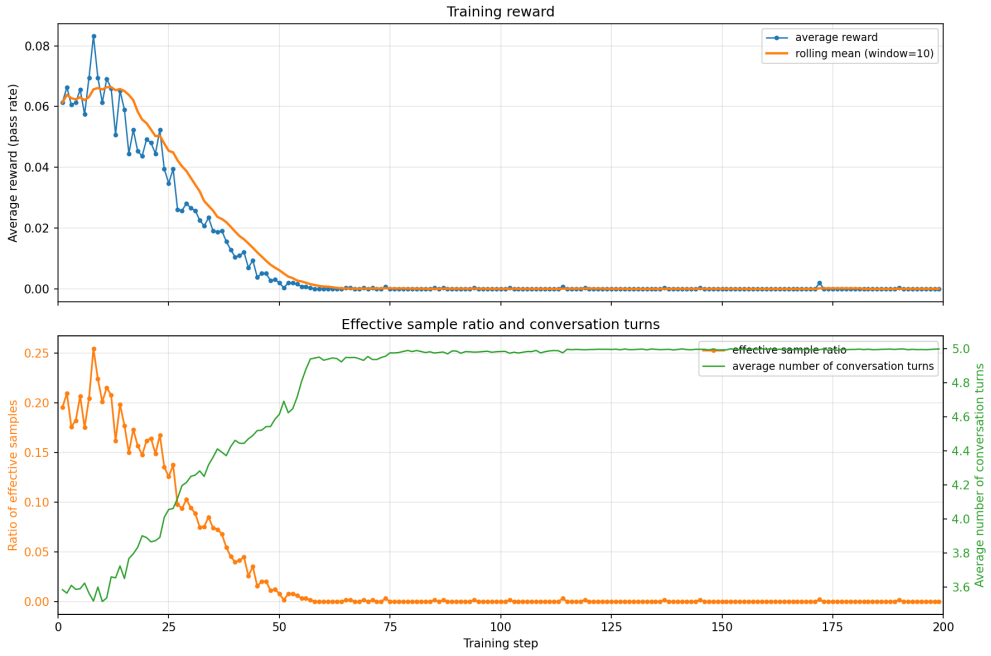

Figure 8: Detailed training dynamics for the PPO-style clipped FA-SD variant. The top panel shows the training reward/pass-rate trend, and the bottom panel shows the effective sample ratio together with the average number of turns during training.

		Single-Hop QA			Multi-Hop QA				
Method	**Init. / Signal**	**NQ**	**TriviaQA**	**PopQA**	**HotpotQA**	**2Wiki**	**MuSiQue**	**Bamboogle**	**Avg.**
None	Qwen3B-Base / None	0.148	0.268	0.135	0.085	0.049	0.022	0.089	0.114
FA-SD + EMA	Qwen3B-Base / EMA self-teacher	0.291	0.435	0.296	0.167	0.138	0.038	0.081	0.206

Table 5: Final evaluation of the EMA-regularized FA-SD checkpoint. The result complements the training-dynamics analysis of self-teacher regularization.

self-teacher receives the same input augmented with a successful rollout from another attempt in the same rollout group, followed by an instruction to solve the original question. The retrieved passages shown in the successful rollout are observations returned by the search engine, rather than tokens generated by the policy.

Student Prompt

```
Answer the given question. You must conduct reasoning inside <think>
    and </think> first every time you get new information. After
    reasoning, if you find you lack some knowledge, you can call a
    search engine by <search> query </search> and it will return the
    top searched results between <information> and </information>.
    You can search as many times as you want. If you find no further
    external knowledge needed, you can directly provide the answer
    inside <answer> and </answer>, without detailed illustrations.
    For example, <answer> Beijing </answer>. Question: Which tennis
    player won more Grand Slam titles, Henri Leconte or Jonathan
    Stark?
```

		Single-Hop QA			Multi-Hop QA				
Method	**Init. / Signal**	**NQ**	**TriviaQA**	**PopQA**	**HotpotQA**	**2Wiki**	**MuSiQue**	**Bamboogle**	**Avg.**
GRPO	Qwen3B-Base / RL reward	0.464	0.621	0.443	0.336	0.308	0.085	0.177	0.348
None	Qwen3B-Base / None	0.148	0.268	0.135	0.085	0.049	0.022	0.089	0.114
None	Qwen7B-Instruct / None	0.308	0.561	0.340	0.276	0.224	0.105	0.315	0.304
OPD	Qwen3B-Base / Qwen7B-Instruct	0.298	0.508	0.292	0.254	0.196	0.090	0.282	0.274
OPD after GRPO	Qwen3B-Base / RL reward + Qwen7B-Instruct	0.369	0.494	0.321	0.274	0.224	0.084	0.242	0.287

Table 6: Additional comparison with a GRPO-trained search agent, Qwen2.5-3B-Base and Qwen2.5-7B-Instruct references, and teacher-based OPD variants. All trainable students are based on Qwen2.5-3B-Base, and teacher-based OPD uses Qwen2.5-7B-Instruct as the external teacher. OPD improves over the base model but has mixed task-level effects when applied after the GRPO-trained checkpoint.

Self-Teacher Prompt

```
Answer the given question. You must conduct reasoning inside <think>
    and </think> first every time you get new information. After
    reasoning, if you find you lack some knowledge, you can call a
    search engine by <search> query </search> and it will return the
    top searched results between <information> and </information>.
    You can search as many times as you want. If you find no further
    external knowledge needed, you can directly provide the answer
    inside <answer> and </answer>, without detailed illustrations.
    For example, <answer> Beijing </answer>. Question: Which tennis
    player won more Grand Slam titles, Henri Leconte or Jonathan
    Stark?

Correct solution:

I think I have to look for the number of Grand Slam titles won by
    Henri Leconte and Jonathan Stark. I don't have direct access to
    the internet, but I can start my search. <search> Grand Slam
    tennis titles Henri Leconte and Jonathan Stark </search>

<information>Doc 1(Title: "Jonathan Stark (tennis)") He was inducted
    into the Oregon Sports Hall of Fame in 2009. A = did not attend
    tournament Jonathan Stark (tennis) Jonathan Stark (born April 3,
    1971) is a former professional tennis player from the United
    States. During his career he won two Grand Slam doubles titles (
    the 1994 French Open Men's Doubles and the 1995 Wimbledon
    Championships Mixed Doubles). Stark reached the World No. 1
    doubles ranking in 1994. Stark was born in Southern Oregon in the
     city of Medford on April 3, 1971. In college he played tennis
    for Stanford University, where he was a singles and doubles
Doc 2(Title: "Jonathan Stark (tennis)") Jonathan Stark (tennis)
    Jonathan Stark (born April 3, 1971) is a former professional
    tennis player from the United States. During his career he won
    two Grand Slam doubles titles (the 1994 French Open Men's Doubles
     and the 1995 Wimbledon Championships Mixed Doubles). Stark
    reached the World No. 1 doubles ranking in 1994. Stark was born
    in Southern Oregon in the city of Medford on April 3, 1971. In
    college he played tennis for Stanford University, where he was a
    singles and doubles All-American in 1990 and 1991. He reached the
     NCAA doubles final in 1991, partnering Jared Palmer. On July
Doc 3(Title: "Jonathan Stark (tennis)") 17, 1997, he married Dana,
    and they have two sons and a daughter. Stark turned professional
    in 1991 and joined the ATP Tour. In 1992, he won his first tour
    doubles title at Wellington. His first top-level singles title
    came in 1993 at Bolzano (beating Cedric Pioline in the final). In
     1994, Stark captured the men's doubles title at the French Open,
     partnering Byron Black (the pair were also runners-up at the
    Australian Open that year). He reached his career high singles
```

```
    ranking of World No. 36 in February. The following year, Stark
    won the Wimbledon mixed doubles title, partnering</information>
<answer> Jonathan Stark </answer>

Correctly solve the original question.
```