Meta-Reinforcement Learning with Self-Reflection for Agentic Search

Teng Xiao[1,2*] **Yige Yuan**[2*] **Hamish Ivison**[1,2] **Huaisheng Zhu**[3] **Faeze Brahman**[1]
Nathan Lambert[1] **Pradeep Dasigi**[1,†] **Noah A. Smith**[1,2,†] **Hannaneh Hajishirzi**[1,2,†]
[1]Allen Institute for AI [2]University of Washington [3]Independent

Abstract

This paper introduces MR-Search, an in-context meta reinforcement learning (RL) formulation for agentic search with self-reflection. Instead of optimizing a policy within a single independent episode with sparse rewards, MR-Search trains a policy that conditions on past episodes and adapts its search strategy across episodes. MR-Search *learns to learn* a search strategy with self-reflection, allowing search agents to improve in-context exploration at test-time. Specifically, MR-Search performs cross-episode exploration by generating explicit self-reflections after each episode and leveraging them as additional context to guide subsequent attempts, thereby promoting more effective exploration during test-time. We further introduce a multi-turn RL algorithm that estimates a dense relative advantage at the turn level, enabling fine-grained credit assignment on each episode. Empirical results across various benchmarks demonstrate the advantages of MR-Search over baselines based RL, showing strong generalization and relative improvements of **9.2**% to **19.3**% across eight QA benchmarks. Our code and data are available at https://github.com/tengxiao1/MR-Search.

1 Introduction

Language models with advanced reasoning capabilities have driven substantial progress toward more autonomous and multi-step decision-making behaviors in complex tasks (Guo et al., 2025; Jaech et al., 2024). Examples include agentic search such as deep research (Du et al., 2025; Shao et al., 2025) and other information seeking (Mialon et al., 2023; Jin et al., 2025a), where LMs use search tools and engage in dynamic, multi-turn interactions. Reinforcement learning (RL) with the ReAct paradigm (Yao et al., 2022) has emerged as a primary framework for training search agents (Jin et al., 2025a; Zheng et al., 2025), replacing the traditional reliance on supervised data collection. Specifically, these methods mainly focus on the correctness of the final answer and only receive sparse rewards at the end of each trajectory, without providing precise credit assignment for intermediate steps. Due to the sparse nature of outcome rewards, the agent often struggles to learn more complex processes and is susceptible to issues such as inefficient exploration at the early stage, local optima, and inefficient search dynamics (Zhang et al., 2025; Feng et al., 2025). These challenges become more pronounced in agentic search tasks, where multi-turn interactions with tools could amplify small errors and obscure credit assignment (Feng et al., 2025).

To address the key challenge of sparse rewards, several works have explored using *process* reward models (Luo et al., 2024; Wang et al., 2023) or LM judges (Zheng et al., 2024; Deng et al., 2025). However, these approaches rely on external annotations, which are both costly and difficult to reuse when task requirements change. Moreover, model-based rewards inevitably lead to reward hacking and bias (Wang et al., 2025a) and incur additional computational overhead in the RL training (Yuan et al., 2024).

In this paper, we introduce MR-Search, a simple yet effective meta-RL approach that enables search agents to improve in-context exploration at test time. Our work is most closely

*Equal contribution. Correspondence to tengx@allenai.org. †Equal senior authors.

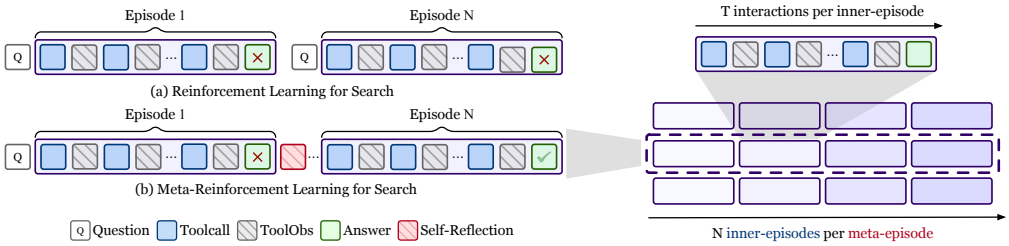

Figure 1: RL-based agents (a) condition solely on the current episode, and episodes are *independent*, whereas meta-RL-based agents (b) leverage context accumulated across episodes. `MR-Search` performs sequential self-reflection over past episodes to guide exploration in subsequent episodes. In `MR-Search`, we have inner-episodes, each consisting of a maximum of T interactions steps with an answer. A sequence of N episodes forms a meta-episode.

related to in-context meta-reinforcement learning methods (Duan et al., 2016; Stadie et al., 2018; Laskin et al., 2023), which leverage in-context histories from a few initial exploration episodes to guide subsequent exploitation episodes. Unlike traditional meta-RL approaches (Duan et al., 2016; Stadie et al., 2018; Laskin et al., 2023) in robotics and games, we focus on open-domain agentic search tasks with tool interactions and self-reflection, without any reward feedback from the environment during inference.

As illustrated in Figure 1, `MR-Search` formulates agentic search as an iterative, self-reflection driven process instead of performing exploration through multiple independent episodes that operate in isolation in RL-base agents. In `MR-Search`, each complete interaction trajectory with an answer is an episode and followed by an explicit reflection step. This design enables sequential self-reflection and cross-episode knowledge consolidation in a multi-turn setting, transforming exploration from a set of disconnected attempts into a progressively informed search process. Thus, the agent learns to balance exploration and exploitation end-to-end by updating its search strategy according to final task performance. Our method can be seen as an instance of meta-learning where we meta-learn how to generate effective self-reflection. `MR-Search` is related to the concept of the Ralph Loop (Carson, 2026), which forces the agent to continue and repeatedly work toward a better solution: *Give a better answer. Don't give up*. Unlike these methods, which operate purely at inference time, `MR-Search` proposes cross-episode meta-RL formulation to enable fast self-reflection during the test time.

To optimize the policy with multi-turn reflection, we use a multi-turn RL algorithm that estimates unbiased grouped relative advantages (Ahmadian et al., 2024) at the turn level to assign localized credit to self-reflection turns. As a result, `MR-Search` remains critic-free and eliminates the need for auxiliary value models compared to PPO (Schulman et al., 2017).

Our main contributions are: (i) We advocate for and formalize in-context meta-reinforcement learning as a practical and scalable bridge between meta-learning and reinforcement learning for agentic search, where ground-truth rewards are absent at inference time. (ii) We propose `MR-Search`, an effective multi-turn agentic search framework that performs cross-episode exploration by generating an explicit self-reflection after each interaction episode. (iii) Empirically, we validate the effectiveness of `MR-Search` across multiple multi-hop QA benchmarks, showing that it significantly outperforms prior methods. Specifically, `MR-Search` achieves an average relative improvement of **9.2**% to **19.3**% over strong baselines.

2 Related Work

RL for Agentic Search. RL has emerged as a promising training paradigm for developing adaptive and autonomous search agents (Jin et al., 2025a; Chen et al., 2025; Wu et al., 2025b; Li et al., 2025a). Specifically, agentic search with RL trains LLMs as decision-making agents that interact with a search through reasoning, receive feedback, and iteratively refine their strategies to maximize task rewards. Despite significant progress, these methods rely solely on sparse outcome rewards, without providing precise credit assignment for effective exploration (Feng et al., 2025). Recently, several works (Deng et al., 2025; Wang et al., 2025b) have attempted to design process rewards for agentic search. However, these approaches

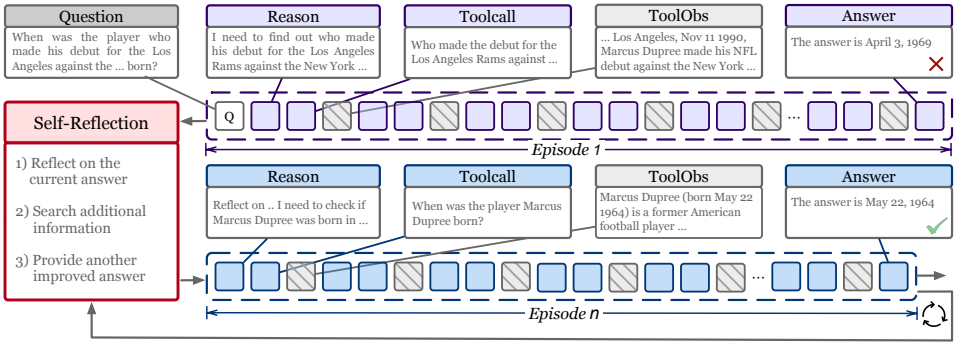

Figure 2: An overview of our proposed `MR-Search` framework. Given a question, the agent first completes an initial episode by interleaving reasoning and tool calls. It then enters an iterative self-reflection loop, where previous episodes serve as experience to inform subsequent searches and answer revisions, enabling iterative improvement across episodes.

require annotations at every intermediate step or rely on external evaluators, both of which are expensive to obtain. In contrast, our `MR-Search` aligns with in-context meta-RL, enabling progressively targeted exploration with explicit cross-episode reflection.

Meta-Reinforcement Learning. Our work is related to meta-RL methods that leverage in-context histories to guide exploration, with the goal of maximizing rewards in subsequent exploitation episodes in robotics and game domains (Melo, 2022; Laskin et al., 2023). Qu et al. (2025) formulate the optimization of test-time compute for LLMs as a meta-reinforcement learning problem. Related work (Jiang et al., 2025) similarly proposed using meta-RL to encourage exploration in LLM agents, with ground-truth feedback available during inference in game environments. In contrast, our work focuses on open-domain agentic tasks with LLMs and assumes no access to environment feedback during inference.

LLMs with Self-Reflection. Previous studies investigate prompting-based strategies that enable LLMs to iteratively and sequentially refine their own generations via intrinsic feedback, effectively scaling test-time computation. These approaches include self-correction (Huang et al., 2023), self-reflection (Shinn et al., 2023), and self-refine (Madaan et al., 2023). Beyond prompting alone, several subsequent works (Wan et al., 2025; Qu et al., 2024; Kumar et al., 2024; Xiong et al., 2025; Yuksekgonul et al., 2026) employ finetuning to train models for intrinsic self-correction. A recent work (Shi et al., 2026) also proposes experiential reinforcement learning to incorporate textual reflection feedback. In contrast, our work offers a novel meta-RL perspective on agentic search with tool interaction to catalyze continuous self-reflection, enabling the model to more effectively explore better answers.

Test-time Scaling. The paradigm of scaling test-time compute has emerged as a critical avenue for enhancing reasoning capabilities (Guo et al., 2025). Prior work explores two main dimensions: parallel sampling (Wang et al., 2022b) and sequential refinement (Madaan et al., 2023; Snell et al., 2024; Wang et al., 2026). Parallel sampling generates multiple answers independently, whereas sequential refinement generates answers sequentially, with each attempt conditioned on previous ones. Our work is most closely related to the sequential refinement paradigm. However, unlike these methods, which operate purely at inference time, this work proposes cross-episode meta-learning to enable continuous self-reflection in context, showing that meta-RL induces effective in-context exploration for agentic search.

3 MR-Search: Meta-Reinforcement Learning for Agentic Search

3.1 Background

Given a dataset $\mathcal{D}$ consisting of question–answer pairs and an external search engine $\mathcal{E}$, our goal is to train a LLM agent to give the answer by iteratively performing reasoning and interacting with $\mathcal{E}$. Many search agent frameworks build upon the ReAct paradigm (Yao et al., 2022; Jin et al., 2025a), where agents execute iterative cycles of reasoning and acting

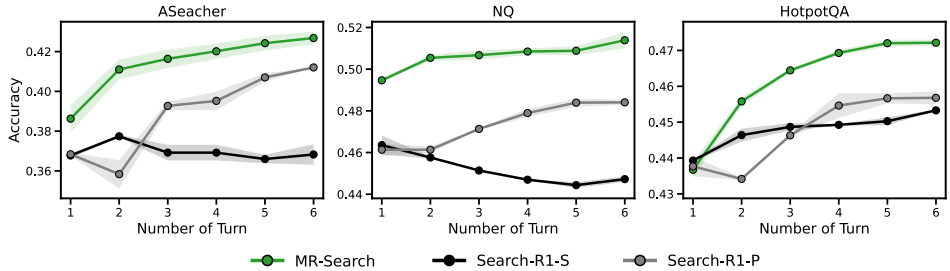

Figure 3: We evaluate MR-Search, Search-R1 with sequential reflection inference (Search-R1-S), and Search-R1 with parallel sampling (Search-R1-P), selecting the most frequent answer among the generated trajectories. Shaded regions show the standard deviation across 3 runs. We observe that MR-Search achieves the best performance. See § 4.3 for details

until a final answer is reached. When presented with a query, the agent conducts multiple rounds of thought-action-observation sequences. During each round, the RL-based agent π_θ formulates a *internal* Thought (τ) based on the current context, executes a *external* Action (α) such as query, and receives corresponding feedback from search engines as tool Observation (x) as shown in Figure 1. The interaction trajectory is denoted as:

$$a = (\tau_0, \alpha_0, x_0, \tau_1, \alpha_1, x_1, \ldots, \tau_{T-1}) \,. \tag{1}$$

The first round contains the prompt, while the final round τ_{T-1} contains only the thought with the final answer o, without any further actions. Given this interaction process, we can directly maximize the RL objective with the final outcome rewards to optimize the policy:

$$\mathcal{J}(\pi_\theta) = \mathbb{E}_{a \sim \pi_\theta}[f_{\text{verifier}}(o, o^*)], \tag{2}$$

where o denotes the final answer extracted from the completed trajectory a, o^* is the ground-truth answer, and f_{verifier} represents either a rule-based or model-based verifier. Although agentic search based on outcome rewards has demonstrated promising performance (Jin et al., 2025a; Sun et al., 2025; Chen et al., 2025; Shi et al., 2025; Wu et al., 2025b), the outcome rewards are sparse and delayed, leading to ambiguous credit assignment and ineffective search exploration (Liu et al., 2021; Feng et al., 2025; Wang et al., 2025b).

3.2 Meta-RL Framework for Agentic Search

In this section, we introduce MR-Search, a Meta-RL framework for agentic search built on a cross-episode meta training scheme. Each meta-episode is modeled as a sequence of episodes, encouraging early exploration and subsequent exploitation of accumulated context, as shown in Figure 1. By modeling prior episodes in context through a standardized self-reflection paradigm and propagating this information across episodes, MR-Search enables increasingly informed search, leading to broader exploration and reduced redundancy.

Algorithm 1 MR-Search: One Sample Update

1: **Input:** policy π_θ, sample $(x, o^*) \sim \mathcal{D}$, group size G, reflection steps N
2: **for** $i = 1, \ldots, G$ **do** ▷ Group of meta-episodes
3: $C \leftarrow x$ ▷ Initialize context
4: **for** $n = 0, \ldots, N-1$ **do** ▷ Inner episodes
5: $a_{i,n} \sim \pi_\theta(\cdot \mid C)$
6: $C \leftarrow C \oplus a_{i,n} \oplus \textsc{Reflect}(C)$ ▷ Self-reflection
7: $r_{i,n} \leftarrow f(a_{i,n}, o^*)$ ▷ Reward
8: Compute $A_{i,n}$ via Eq. (7–8) ▷ RLOO advantage
9: $\theta \leftarrow \text{RL_}\textsc{Update}(\theta, \{r_{i,n}\})$ with Eq. (9)

The key motivations are twofold: (i) For LLM agents, model limitations often stem from insufficient exploration within a single trajectory rather than inadequate reasoning capacity (Shen et al., 2025); incorrect answers may consistently appear across multiple samples in parallel, as shown by Fan et al. (2025); Si et al. (2026). (ii) LLMs exhibit strong self-reflection capabilities, enabling inference-time adaptation that accumulates information from prior interactions across episodes, naturally leveraging in-context learning mechanisms. As shown in Figure 3, MR-Search substantially improves over baselines both with sequential reflection

and parallel sampling and its performance grows with more turns, suggesting that the ability for self-reflection emerges as models become stronger in-context meta-learners.

Given an input question, the search agent π_θ first generates the first episode, in which it interacts with external tools until reaching the final answer, following the procedure in §3.1.

$$a_0 \sim \pi_\theta(a). \tag{3}$$

Unlike standard RL-based search agents (Jin et al., 2025a; Song et al., 2025), where episodes are independent, `MR-Search` conditions each episode on the preceding one. After each episode, the search agent invokes the tool to iteratively improve its answer under the self-reflection paradigm. Specifically, given an initial episode, we apply a reflection prompt (Appendix A.1.3) that triggers the model to refine its answer by conditioning on previous episodes as context and producing an additional episode with another answer.

$$a_1 \sim p_\theta(a_1 \mid a_0), \;\; a_2 \sim p_\theta(a_2 \mid a_0, a_1) \ldots \tag{4}$$

By repeating the multi-turn self-reflection process $N-1$ times, in `MR-Search`, each meta-episode consists of N episodes sequentially generated by the search agent:

$$y = (a_0, a_1, \cdots, a_N), \tag{5}$$

where we can compute rewards independently for each episode a using the answer y_n extracted from the nth episode by the verifier. Given this multi-turn process, we can define the meta-level objective as maximizing the expected reward of the meta-episode:

$$\mathcal{J}_{\text{meta}}(\pi_\theta) = \mathbb{E}_{y\sim\pi_\theta}\Big[\sum_{n=0}^{N-1}\gamma^n R(s_n, a_n)\Big] = \mathbb{E}_{y\sim\pi_\theta}\Big[\sum_{n=0}^{N-1}\gamma^n f_{\text{verifier}}(o_n, o^*)\Big], \tag{6}$$

where $s_n = a_{<n}$ denotes the accumulated meta context up to episode n, o_n is the answer extracted from the episode a_n, and $\gamma \in (0, 1]$ is the discount factor accounting for future returns. Unless otherwise specified, we set $\gamma = 1$ in this work. In this work, `MR-Search` conditions on trajectories and explicit reflections from all previous episodes, causing the context length to increase linearly with the number of reflection steps N. To mitigate this scalability issue, one can retain only the immediately preceding episode as context, or adopt a context management protocol that summarizes prior episodes before carrying them forward to subsequent episodes. In §4.4, we empirically find that only keeping the immediately preceding episode as context also works well for `MR-Search`.

3.3 Policy Optimization with Multi-Turn Advantages

To optimize the policy, instead of estimating advantages using separate value functions as in PPO (Schulman et al., 2017), we propose an approach that incorporates turn-level reward signals while maintaining unbiased policy optimization. Specifically, for each question, we sample a group of G meta-episodes, $\mathcal{G} = \{y_i\}_{i=1}^G$. To make rewards comparable across episodes and reduce variance, we aggregate rewards at the same episode over meta-episodes and apply Leave-One-Out (RLOO) estimation (Kool et al., 2019; Ahmadian et al., 2024):

$$\tilde{r}_{i,n} = r(s_{i,n}, a_{i,n}) - \text{mean}_{j\neq i} r(s_{j,n}, a_{j,n}) = r(s_{i,n}, a_{i,n}) - \frac{1}{G-1}\sum_{j\neq i} r(s_{j,n}, a_{j,n}), \tag{7}$$

where $\tilde{r}_{i,n}$ is the reward of the nth episode in the ith meta-episode (i.e., $y_{i,n}$). Compared to GRPO (Shao et al., 2024), RLOO provides an unbiased advantage estimation (Bereket & Leskovec, 2025). While $\tilde{r}_{i,n}$ captures the relative quality of each episode, it reflects immediate effects and ignores the impact of future. To include long-horizon dependencies, we compute a discounted cumulative advantage to propagate rewards backward to earlier turns:

$$A_{i,n} = \sum_{n'=n}^{N} \gamma^{n'-n}\tilde{r}_{i,n'}. \tag{8}$$

As the baseline used in RLOO, $\text{mean}_{j\neq i} r(s_{j,n}, s_{j,n})$, does not depend on the current action $y_{i,n}$, it provides an unbiased estimate of the turn-level advantage.

With the discounted turn-level advantages above, we optimize the policy using a clipped surrogate off-policy objective in PPO (Schulman et al., 2017). Formally, the objective is:

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{n=1}^{|y_i|}\min\Big(\frac{\pi\left(y_{i,n} \mid x, y_{i,<n}; \theta\right)}{\pi(y_{i,n} \mid x, y_{i,<t}; \overline{\theta})} A_{i,n},\; \text{clip}\big(\frac{\pi\left(y_{i,n} \mid x, y_{i,<n}; \theta\right)}{\pi(y_{i,n} \mid x, y_{i,<n}; \overline{\theta})},\; 1-\varepsilon,\; 1+\varepsilon\big)\, A_{i,n}\Big), \tag{9}$$

where $\pi(y_{i,n} \mid x, y_{i,<n}; \theta)$ and $\pi(y_{i,n} \mid x, y_{i,<n}; \bar{\theta})$ denote the current and old policy models over the steps, and we broadcast each step's advantage signal to all tokens in that step (Shao et al., 2024). ϵ is the is the clipping ratio. We additionally mask out tool output tokens from the loss, following previous work (Jin et al., 2025a). Optimizing the policy with the above objective enables the policy to capture both global trajectory quality and local step effectiveness. In §4.2, we compare our objective with PPO (Schulman et al., 2017) and MT-GRPO (Zeng et al., 2025) using our designed process rewards. The results show that our objective consistently achieves better performance under the meta-RL framework. The full algorithm of `MR-Search` (one-sample training) is summarized in Algorithm 1.

3.4 Discussion

Exploration & Exploitation. As discussed above, `MR-Search` leverages experience from previous episodes to guide subsequent exploration. By default, all episodes contribute rewards according to Eq. (8). To promote unstructured exploration, we can optionally mask rewards for designated exploration episodes while retaining rewards for exploitation episodes (Stadie et al., 2018). Exploration episodes are fully visible during the forward pass but receive zero reward during backpropagation, so gradients are driven only by exploitation episodes. Concretely, the advantage is computed using a masked return:

$$A_{i,n} = \sum_{n'=n}^{N} \gamma^{n'-n} \tilde{r}_{i,n'} m_{n'}, \tag{10}$$

where $m_{n'} \in \{0, 1\}$ indicates exploitation (1) or exploration (0). The policy gradient is therefore computed using this masked return rather than the standard discounted return. By zeroing out exploration rewards, we encourage the policy to prioritize long-term gains from improved context adaptation rather than short-term episode feedback. Although exploration episodes do not directly contribute to the gradient, they serve as contextual adaptation steps that improve environment identification and lead to higher rewards in subsequent exploitation episodes. We provide empirical analysis of this strategy in §4.4 and find that it is helpful for ASearcher, which requires multi-turn search.

Meta-RL at the Step Level. While `MR-Search` models agentic search as a meta-episode composed of multiple reflection episodes, it treats each full trajectory with an answer as a single optimization unit. However, in long-horizon reasoning, inefficiencies often arise at a finer granularity, such as individual tool calls or intermediate reasoning steps. The same principle of `MR-Search` extends naturally to semantically meaningful sub-episodes. For instance, in agentic search, each tool-interaction step can be treated as a micro-episode. During training, we prompt the model to produce an intermediate answer after each tool call as shown in Figure 8 in Appendix and evaluate it with the verifier to obtain step-level rewards. Converting these substructures into micro-episodes enables localized credit assignment, transforming long trajectories into reflection steps. This dense supervision promotes informative intermediate reasoning and reduces redundant exploration. We empirically show that this extension also achieves strong performance (§4.4).

4 Experiments

4.1 Experimental Setup

Datasets We conduct evaluations on the following datasets: *(1) General Question Answering*: NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2022); and *(2) Multi-Hop Question Answering*: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), Musique (Trivedi et al., 2022), and Bamboogle (Press et al., 2022). For training, we merge the NQ and HotpotQA training sets to construct a unified dataset for all finetuning approaches, following the setup in (Jin et al., 2025b). For evaluation, we use the test or development splits of the seven datasets listed above. We additionally include a synthetic dataset, ASearcher (Gao et al., 2025), which is more complex than NQ/HotpotQA and requires long multi-turn search. We split ASearcher into 90% training and 10% evaluation sets. Detailed dataset descriptions and statistics are provided in Appendix A.1.1.

Table 1: Main accuracy (%) on search-based QA benchmarks. The best results are marked in **boldface**. We compare with baselines that rely on outcome rewards (ReSearch and Search-R1) and those that use process rewards with external models (PPRM and StepResearch).

	Method	Single-Hop QA			Multi-Hop QA				
		NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	Avg.
Qwen2.5-3b	Direct Inference	10.6	28.8	10.8	14.9	24.4	2.0	2.4	13.4
	Search-o1	23.8	47.2	26.2	22.1	21.8	5.4	32.0	25.5
	ReSearch	42.7	59.7	43.0	30.5	27.2	7.4	11.5	30.4
	Search-R1	46.2	62.2	45.6	32.6	31.0	7.7	17.6	34.7
	PPRM	42.3	56.5	41.1	35.3	34.0	12.7	28.0	35.7
	StepResearch	44.6	61.5	45.6	37.3	33.8	10.5	32.5	38.0
	`MR-Search`	**47.7**	**63.5**	**46.0**	**41.9**	**40.1**	**16.5**	**34.4**	**41.4**
Qwen2.5-7b	Direct Inference	13.4	40.8	14.0	18.3	25.0	3.1	12.0	18.1
	Search-o1	15.1	44.3	13.1	18.7	17.6	5.8	29.6	20.6
	ReSearch	36.6	60.5	39.1	37.8	38.6	16.6	37.6	38.1
	Search-R1	45.9	63.2	44.9	43.9	38.7	18.1	40.0	42.1
	PPRM	45.8	61.0	43.7	38.6	35.5	14.7	35.5	39.3
	StepResearch	47.3	63.6	43.1	43.9	41.8	20.5	43.5	43.4
	`MR-Search`	**50.2**	**66.6**	**47.2**	**46.8**	**43.6**	**22.1**	**45.2**	**46.0**

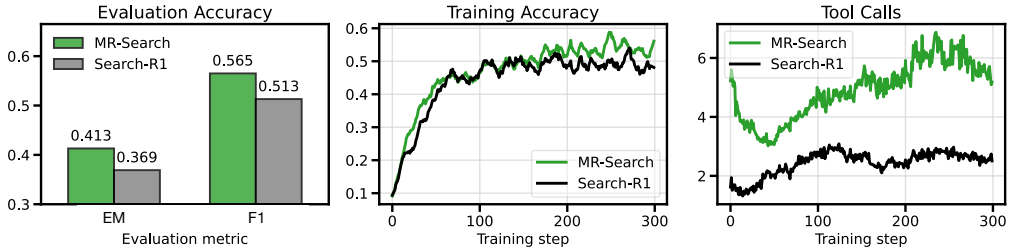

Figure 4: Test performance, training curves of reward and search frequency on ASearcher, evaluated with Qwen2.5-7B-Base. Additional results are provided in Appendix A.2.1.

Evaluation Metrics For evaluation metrics, we follow Jin et al. (2025b): we first normalize both the predicted and ground-truth answers, and then compute Exact Match (EM) score. EM achieves true if and only if the predicted answer exactly matches any ground-truth answer. For all methods, we sample a single trajectory per question and report the average EM for the last valid prediction for questions following Jin et al. (2025a).

Setup We mainly conduct experiments using Qwen-series models (Yang et al., 2024) (Qwen2.5-3B-Base and Qwen2.5-7B-Base). Following (Jin et al., 2025b), we use the 2018 Wikipedia dump (Karpukhin et al., 2020) as the knowledge source and E5 (Wang et al., 2022a) as the retriever. We fix the number of retrieved documents to three across all methods for a fair comparison. Unless otherwise specified, we report results at turn 3 for our models. The detailed implementation settings are provided in Appendix A.1.3.

Baselines To evaluate the effectiveness of `MR-Search`, we compare it against the following recent baselines: (1) Inference without finetuning: methods that directly use the base model, including direct inference without retrieval and Search-o1 (Li et al., 2025b) with retrieval. (2) Finetuning-based methods that learn a policy to integrate the search tool without step-level supervision, including ReSearch (Chen et al., 2025) and Search-R1 (Jin et al., 2025a). (3) Finetuning-based methods that learn a policy to integrate the search tool with step-level supervision, including PPRM (Anonymous, 2026) and StepResearch (Wang et al., 2025b). We also compare with Search-R1 with our multi-turn reflection mechanism during inference. in § 4.3. The detailed description of baselines are given in Appendix A.1.2.

4.2 Main Results

Table 1 summarizes main results on benchmarks. Among the approaches, our `MR-Search` achieves a substantial margin over GRPO with outcome rewards (Search-R1), yielding 9.2%

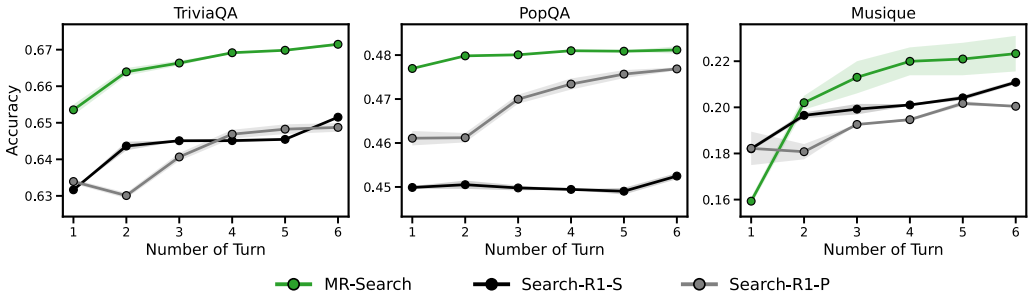

Figure 5: MR-Search, Search-R1 with sequential reflection turns (Search-R1-S), and Search-R1 with parallel sampling (Search-R1-P), selecting the most frequent answer among them.

Table 2: Ablation of discount factor and training algorithm evaluated on Qwen2.5-7B-Base.

Method	Single-Hop QA			Multi-Hop QA				Avg.
	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	
ReSearch	36.6	60.5	39.1	37.8	38.6	16.6	37.6	38.1
Search-R1	46.4	64.1	44.8	43.0	42.5	19.5	44.0	43.5
MR-Search w. $\gamma = 0$	49.3	63.9	45.6	44.3	41.7	19.3	42.6	43.8
MR-Search w. PPO	43.9	63.7	42.5	41.3	40.5	18.6	43.5	42.0
MR-Search w. MT-GRPO	46.1	64.5	44.5	44.7	**44.2**	20.7	45.1	44.3
MR-Search	**50.2**	**66.6**	**47.2**	**46.8**	43.6	**22.1**	**45.2**	**46.0**

and 19.3% relative improvements on average for the Qwen2.5-7B-Base and Qwen2.5-3B-Base, respectively. This highlights the significant benefits of designing process rewards for agentic search. Remarkably, MR-Search remains highly effective on the small Qwen2.5-3B model, whereas RL methods that rely only on sparse outcome rewards struggle to elicit multi-turn search behavior for good performance. Compared to other methods that rely on external models to obtain the process reward such as StepResearch and PPRM, our MR-Search achieves better performance. This confirms that agentic search can benefit greatly from our designed free process rewards, and that MR-Search can effectively leverage this process supervision to achieve better performance. Figure 4 shows the results on the ASearcher datasets. Compared to Multi-Hop QA, ASearcher requires longer-horizon, multi-turn search (Gao et al., 2025). From the results, we observe that MR-Search significantly outperforms Search-R1, achieving 10.2% and 9.5% relative improvements in EM and F1, respectively, demonstrating the effectiveness of MR-Search on complex ASearcher tasks.

Ablation Study. We study the effects of the key design choices of MR-Search. Specifically, we consider the following ablations: (i) we compare our optimization algorithm (Section §3.3) with PPO (Schulman et al., 2017) and MT-GRPO (Zeng et al., 2025) within our Meta-RL framework; (ii) we set the discount factor $\gamma = 0$, which removes future credit assignment. From the results in Table 2, we observe that our proposed multi-turn RL algorithm consistently outperforms PPO and MT-GRPO when trained with episode turn-level rewards, demonstrating the effectiveness of our optimization strategy in leveraging dense feedback from reflection. Moreover, both PPO and MT-GRPO underperform GRPO with outcome rewards (Search-R1) on single-hop NQ and PopQA, whereas our method does not, indicating stronger generalization and robustness. We further observe that removing the discount factor substantially degrades performance and causes the training process to converge to poor local optima. As discussed above, a possible explanation is that an incorrect episode does not necessarily imply that intermediate episodes are uninformative.

4.3 Further Analysis

Training Dynamics. For a more comprehensive understanding of MR-Search, we visualize its training dynamics. Figure 4 shows the training reward dynamics and of Search-R1 and MR-Search. We observe that MR-Search exhibits stable convergence during training and consistently achieves higher training reward than Search-R1. This indicates that MR-Search effectively leverages iterative reflection to refine answers, leading to progressively improved final answers throughout training of multi-turn RL training. We can observe that MR-Search

Table 3: Comparison with variants of MR-Search: encouraging exploration, MR-Search at the step level and context management (keeping one preceding episode as context).

Method	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	ASearcher
Search-R1	45.9	63.2	44.9	43.9	38.7	18.1	40.0	36.9
MR-Search	**50.2**	**66.6**	**47.2**	**46.8**	**43.6**	**22.1**	**45.2**	41.3
MR-Search Exploration	48.3	65.1	46.4	44.7	39.4	21.8	44.0	**43.2**
MR-Search Step Level	48.6	64.6	45.7	42.3	41.4	16.3	41.6	38.4
MR-Search Short Context	48.1	65.9	45.2	44.6	41.0	19.3	47.2	40.5

calls the search engine more frequently than Search-R1, demonstrating that MR-Search can dynamically adjust the number of search calls according to the complexity of tasks.

Test-time Scaling. We also evaluate how MR-Search scales with additional reflection turns at test time. We extrapolate the number of reflection turns beyond training (3 turns) by appending the entire interaction history to the context at each turn. As shown in Figures 3 and 5, the single-turn method Search-R1 with our reflection mechanism yields only marginal gains when additional reflection turns are allowed, since its training objective is optimized for a single turn. However, when additional reflection turns are allowed, MR-Search achieves significantly higher performance, exhibiting the steep improvement curve. These results suggest that multi-turn reflection with MR-Search enhances the model's ability to iteratively refine and optimize its search across turns and enables effective extrapolation.

Case Study. We present inference cases of models trained with MR-Search. As illustrated by the examples in Appendix A.2.2, the model is able to execute multi-turn agentic tasks through iterative tool calls and autonomous information aggregation. During subsequent reasoning episode, the model revisits and revises answers in light of newly acquired evidence, ultimately producing correct final answers. These observations highlight the benefits of self-reflection for agentic search. Moreover, when the answer is already accurate as shown in Cases 2 and 3, the model can preserve it and avoid unnecessary revisions, demonstrating its ability to selectively refine reasoning based on retrieved information.

4.4 Extensions

In this section, we conduct preliminary experiments to investigate the effects of exploration vs. exploitation, step-level meta-RL (discussed in § 3.4), and context management (i.e., retaining only the immediately preceding episode as context). To study exploration and exploitation, we designate the first two episodes as exploration and the last two as exploitation. Table 3 in provides detailed results. From Table 3, we observe that all variants significantly outperform the baseline Search-R1, demonstrating the effectiveness of these variants. Moreover, MR-Search with step-level meta-RL achieves a substantial improvement over GRPO with outcome rewards (Search-R1). This result suggests that agentic search can benefit from process-level rewards under meta-RL training, and that MR-Search effectively leverages such process supervision to achieve better performance. Furthermore, as shown in Table 3, encouraging exploration by assigning rewards to the first episode is beneficial for the more complex ASearcher dataset, which requires more interaction with tools.

5 Conclusions

We study agentic search under sparse outcome rewards and propose MR-Search, an in-context meta-reinforcement learning framework that enables structured cross-episode exploration via explicit self-reflection. By conditioning each episode on prior trajectories and reflections, MR-Search transforms independent search attempts into a progressively informed search process, improving exploration without relying on external process reward models. To support this multi-turn reflective setting, we introduce a turn-level grouped advantage formulation that provides unbiased and fine-grained credit assignment while remaining critic-free. Extensive experiments across diverse benchmarks show that MR-Search consistently outperforms outcome-only RL baselines. Overall, our results highlight the importance of in-context meta-learning for effective agentic reinforcement learning of LLMs.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Award No. 2413244. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Ethics Statement

This paper advances reinforcement learning for agentic search by improving training efficiency and credit assignment under sparse rewards. By enabling more effective training of language-model-based agents, this work may benefit practical applications that require efficient and reliable reasoning. This research shares the societal implications of machine learning systems more broadly and does not introduce additional ethical concerns beyond those commonly associated with large language models and reinforcement learning.

Reproducibility Statement

We support reproducibility by providing the code and data for this work and by describing the experimental setup in detail. The paper and appendix specify the datasets and splits, prompts, evaluation protocol, retriever setting, model backbones, key hyperparameters, and compute resources used in our experiments.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Anonymous. Principle process reward for search agents. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2026.

Michael Bereket and Jure Leskovec. Uncalibrated reasoning: Grpo induces overconfidence for stochastic outcomes. *arXiv preprint arXiv:2508.11800*, 2025.

Carson. Ralph: An autonomous ai agent loop that runs repeatedly until all prd items are complete. https://github.com/snarktank/ralph, 2026. GitHub repository, accessed 2026-03-30.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, et al. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.

Yong Deng, Guoqing Wang, Zhenzhe Ying, Xiaofeng Wu, Jinzhen Lin, Wenwen Xiong, Yuqin Dai, Shuo Yang, Zhanwei Zhang, Qiwen Wang, et al. Atom-searcher: Enhancing agentic deep research via fine-grained atomic thought reward. *arXiv preprint arXiv:2508.12800*, 2025.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Yuchen Fan, Kaiyan Zhang, Heng Zhou, Yuxin Zuo, Yanxu Chen, Yu Fu, Xinwei Long, Xuekai Zhu, Che Jiang, Yuchen Zhang, et al. Ssrl: Self-search reinforcement learning. *arXiv preprint arXiv:2508.10874*, 2025.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*, 2025.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Yulun Jiang, Liangze Jiang, Damien Teney, Michael Moor, and Maria Brbic. Meta-rl induces exploration in language agents. *arXiv e-prints*, pp. arXiv–2512, 2025.

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*, 2025a.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=Rwhi91ideu.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pp. 6769–6781, 2020.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! *ICLR 2019 workshop: Deep RL Meets Structured Prediction*, 2019.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

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

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Stenberg Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. In *The Eleventh International Conference on Learning Representations*, 2023.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, et al. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*, 2025a.

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

Evan Z Liu, Aditi Raghunathan, Percy Liang, and Chelsea Finn. Decoupling exploration and exploitation for meta-reinforcement learning without sacrifices. In *International conference on machine learning*, pp. 6925–6935. PMLR, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 7, 2022.

Luckeciano C Melo. Transformers are meta-reinforcement learners. In *international conference on machine learning*, pp. 15340–15359. PMLR, 2022.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. *Advances in Neural Information Processing Systems*, 37:55249–55285, 2024.

Yuxiao Qu, Matthew YR Yang, Amrith Setlur, Lewis Tunstall, Edward Emanuel Beeching, Ruslan Salakhutdinov, and Aviral Kumar. Optimizing test-time compute via meta reinforcement fine-tuning. *arXiv preprint arXiv:2503.07572*, 2025.

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

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

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

Junhong Shen, Hao Bai, Lunjun Zhang, Yifei Zhou, Amrith Setlur, Shengbang Tong, Diego Caples, Nan Jiang, Tong Zhang, Ameet Talwalkar, et al. Thinking vs. doing: Agents that reason by scaling test-time interaction. *arXiv preprint arXiv:2506.07976*, 2025.

Taiwei Shi, Sihao Chen, Bowen Jiang, Linxin Song, Longqi Yang, and Jieyu Zhao. Experiential reinforcement learning. *arXiv preprint arXiv:2602.13949*, 2026.

Yaorui Shi, Shihan Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. Search and refine during think: Autonomous retrieval-augmented reasoning of llms. *NeurIPS*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Chenglei Si, Zitong Yang, Yejin Choi, Emmanuel Candès, Diyi Yang, and Tatsunori Hashimoto. Towards execution-grounded automated ai research. *arXiv preprint arXiv:2601.14525*, 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

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

Bradly C Stadie, Ge Yang, Rein Houthooft, Xi Chen, Yan Duan, Yuhuai Wu, Pieter Abbeel, and Ilya Sutskever. Some considerations on learning to explore via meta-reinforcement learning. *arXiv preprint arXiv:1803.01118*, 2018.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. Zerosearch: Incentivize the search capability of llms without searching. *arXiv preprint arXiv:2505.04588*, 2025.

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

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*, 2025.

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*, 2022a.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023.

Teng Wang, Zhangyi Jiang, Zhenqi He, Shenyang Tong, Wenhan Yang, Yanan Zheng, Zeyu Li, Zifan He, and Hailei Gong. Towards hierarchical multi-step reward models for enhanced reasoning in large language models. *arXiv preprint arXiv:2503.13551*, 2025a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Yike Wang, Huaisheng Zhu, Zhengyu Hu, Yige Yuan, Zhengyu Chen, Shakti Senthil, Hannaneh Hajishirzi, Yulia Tsvetkov, Pradeep Dasigi, and Teng Xiao. Rethinking the evaluation of harness evolution for agents. In *COLM 2026 The 2nd Workshop on Lifelong Agents: Learning, Aligning, and Evolving*, 2026.

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

Jiahao Wu, Zhongwen Xu, Qiang Fu, and Wei Yang. Buy 4 reinforce samples, get a baseline for free! *Cut the Bill, Keep the Turns: Affordable Multi-Turn Search RL*, 2025a.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, et al. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025b.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. Self-rewarding correction for mathematical reasoning. *arXiv preprint arXiv:2502.19613*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*, 2024.

Mert Yuksekgonul, Daniel Koceja, Xinhao Li, Federico Bianchi, Jed McCaleb, Xiaolong Wang, Jan Kautz, Yejin Choi, James Zou, Carlos Guestrin, et al. Learning to discover at test time. *arXiv preprint arXiv:2601.16175*, 2026.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. *arXiv preprint arXiv:2505.11821*, 2025.

Kai Zhang, Xiangchao Chen, Bo Liu, Tianci Xue, Zeyi Liao, Zhihan Liu, Xiyao Wang, Yuting Ning, Zhaorun Chen, Xiaohan Fu, et al. Agent learning via early experience. *arXiv preprint arXiv:2510.08558*, 2025.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*, 2024.

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

Generation Prompt Answer the given question. You must conduct reasoning inside `<think>` and `</think>` first every time you get new information. After reasoning, you must call a search engine by `<search>` query `</search>`, and it will return the top search results between `<information>` and `</information>`. After every time you get new information, you must try to provide the answer inside `<answer>` and `</answer>` without detailed illustrations. For example, `<answer>` xxxx `</answer>`. Question: {question}

A Appendix

A.1 Experimental Details

A.1.1 The Details of Datasets

NQ (Kwiatkowski et al., 2019): NQ is question answering (QA) data set. Questions consist of real anonymized, aggregated queries issued to the Google search engine. The training and test sets contain 79,168 and 3,610 samples, respectively.

TriviaQA (Joshi et al., 2017): TriviaQA is a challenging reading comprehension dataset containing over question-answer-evidence triples, which provide high quality distant supervision for answering the questions. The test sets we used contain 11,313 samples.

PopQA (Mallen et al., 2022): PopQA is a large-scale open-domain question answering dataset, consisting of 14,267 entity-centric QA pairs. Each question is created by converting a knowledge tuple retrieved from Wikidata using a template.

HotpotQA (Yang et al., 2018): HotpotQA is a large-scale multi-hop QA benchmark featuring Wikipedia-based Q&A pairs with sentence-level supporting evidence. The training and test sets contain 90,447 and 7,405 samples, respectively.

2WikiMultiHopQA (Ho et al., 2020): 2WikiMultiHopQA is a multi-hop question answering dataset designed to more reliably test a model's inference across multiple pieces of evidence. The test set contains 7,405 QA pairs.

Musique (Trivedi et al., 2022): MuSique is a multihop question answering dataset constructed to enforce genuine multi-step reasoning. The test set contains 2,417 QA pairs.

Bamboogle (Press et al., 2022). Bamboogle is a dataset with multi-hop questions, where all questions are sufficiently difficult to be unanswerable by a popular internet search engine, but where both supporting pieces of evidence can be found in Wikipedia. Bamboogle contains 125 test QA pairs.

ASearcher (Gao et al., 2025). ASearcher is a synthetic multi-turn dataset whose synthesis pipeline is largely based on Wikipedia. We use the preprocessed version from (Wu et al., 2025a), which applies a three-step filtering pipeline, including the removal of Chinese and math samples as well as rejection sampling, resulting in 14k samples. We split ASearcher into 90% training and 10% evaluation sets to decouple the effects of data distribution.

A.1.2 The Details of Baselines

We provide detailed descriptions of baselines. We consider three types of baselines: Inference without fine-tuning, Fine-tuning-based methods without step-level supervision and Fine-tuning-based methods with step-level supervision.

Search-o1 (Li et al., 2025b) is the search-enhanced reasoning framework, which integrates the agentic RAG mechanism and reason-in-document module.

Reflection Prompt Reflect on your current answer to the question and provide another answer by searching for additional external information using search engines. 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 your want. If you find no further external knowledge needed, you can directly provide the answer inside `<answer>` and `</answer>`., without detailed illustrations.

ReSearch (Chen et al., 2025) is an RL-based framework that trains LLMs to interleave reasoning with explicit search actions, deciding when and how to query and then using retrieved evidence to continue multi-hop reasoning.

Search-R1 (Jin et al., 2025a) extends RL-based reasoning by enabling LLMs to autonomously generate search queries during multi-turn reasoning.

PPRM (Anonymous, 2026) is a principle process reward model that provides step-wise signals to guide GRPO-based RL.

StepResearch (Wang et al., 2025b) trains search agents with step-wise PPO using intermediate rewards and token-level supervision to better guide multi-hop retrieval and reasoning.

A.1.3 The Details of Implementation

Our implementation is built upon Search-R1 based on VeRL. The generation prompt and reflection Prompt are given in the colored box.

Hyperparameters. For training, we use AdamW (Loshchilov & Hutter, 2017) as the optimizer and set the learning rate to 1e-6 without warmup. The top-p and temperature for rollout are both set to 1. The total number of training steps is 300. The number of documents returned by the retrieval engine is 3. The group size for advantage calculation is set to 5. The context length is set to 8K and 16K for the NQ/HotpotQA and Asearcher datasets, respectively. We also set the maximum number of tool calls in each episode to 3 and 5 for the NQ/HotpotQA and Asearcher datasets, respectively. For evaluation, all other settings (e.g., the retriever configuration) are kept the same as in training; the only difference is that we disable sampling and use greedy decoding with `temperature` $= 0$, and `top_p` $= 1.0$.

Computation Resources All RL training is performed on 8× NVIDIA Tesla H100 (80GB) GPUs, and we use an additional 2× NVIDIA Tesla H100 (80GB) GPUs to serve the retriever.

A.2 Additional Experimental Results

A.2.1 More Results of Training Dynamics

In Figures 6 and 7, we further visualize the training dynamics of `MR-Search`. We can observe that `MR-Search` exhibits more stable convergence during training and consistently achieves higher reward than Search-R1 baseline. These results indicate that the process rewards at episode level with self-reflection introduced by `MR-Search` provide more informative and reliable learning signals, leading to improved stability and training effectiveness.

A.2.2 Case Study

In Tables 4-6, we provide some inference trajectory of the model trained by `MR-Search`.

Figure 6: The training dynamics of `MR-Search` and Search-R1 in terms of test accuracy, training accuracy, and the number of tool calls on Qwen2.5-3B-Base.

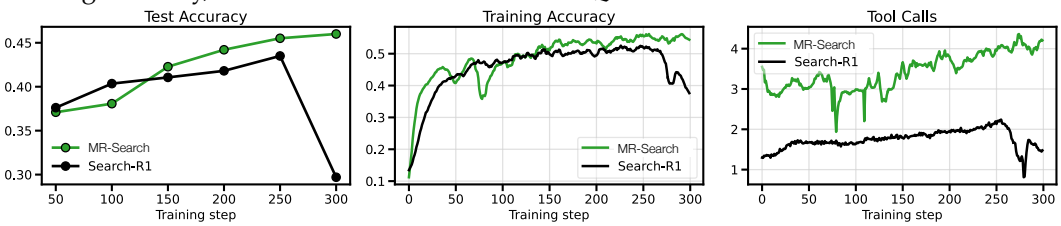

Figure 7: The training dynamics of `MR-Search` and Search-R1 in terms of test accuracy, training accuracy, and the number of tool calls on Qwen2.5-7B-Base.

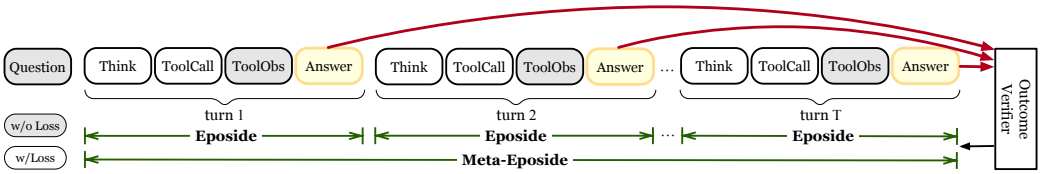

Figure 8: The idea of `MR-Search` can be extended beyond full trajectories to any semantically meaningful episode that can be explicitly segmented. For example, in agentic search, each tool-interaction step can naturally form an episode. We can decompose the one interaction process into multiple self-reflective trials, where each tool-interaction step contributes a candidate answer to the original query via prompting. During training, at each interaction step, we monitor changes in the agent's accuracy with respect to the target and use this signal as a dense training objective to promote effective exploration.

Table 4: Case study 1: An early intermediate answer is incorrect due to ambiguous retrieval results, but is subsequently reconsidered through additional search and self-reflection, leading to the correct final answer.

Question: During a major geopolitical event in the early Cold War, Robert Menzies exploited rising tensions to portray the Labor Party as 'soft on Communism,' leading to the defeat of a Prime Minister who attended a religious school. In the same period, which British maritime patrol aircraft, known for its role in anti-submarine warfare and derived from the world's first jet airliner, was frequently deployed to a U.S. military base in the Atlantic to support anti-submarine operations during the latter stages of the Cold War?

Ground Truth: 2002

MR-Search:
<think> To determine the year in which the Victorian government department responsible for agriculture and energy policy was established, we need more information about the department and its merger with another department in 2013. </think>
<search> Query: Victorian government department responsible for agriculture and energy policy merged with another department administering conservation strategy involving biodiversity offsets in 2013. When was this department established? </search>
<information> Document (Title: Department of the Environment (Australia, 2013–16)) by way of an Administrative Arrangements Order issued on 18 September 2013. It absorbed the responsibilities of the former Department of Sustainability, Environment, Water, Population and Communities (DSWEPaC) and climate change from the former Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education. The department was dissolved in July 2016 and its functions, along with energy policy functions, were moved to the newly established Department of the Environment and Energy. From when it... </information>
<think> It seems the search did not directly provide the information. Let's try another search with a specific focus on the department related to agriculture and energy policy: </think>
<search> Query: When was the Department of Economic Development, Jobs, Transport and Resources established in Victoria? </search>
<information> Document (Title: Department of Economic Development, Jobs, Transport and Resources) Department of Economic Development, Jobs, Transport and Resources The Department of Economic Development, Jobs, Transport and Resources (DEDJTR) is a department of the Government of Victoria. It was created on 1 January 2015 by the government of Premier Daniel Andrews when the number of government departments was reduced from 9 to 7, and assumed responsibility for ministerial portfolios previously spread across 5 departments. The department was established after the 2014 Victorian election by the new Labor government, using the powers of the Premier of Victoria and the Governor-in-council under the Public ... </information>
<think> This information indicates that the Department of Economic Development, Jobs, Transport and Resources was established on 1 January 2015. Therefore, the year in which the Victorian government department responsible for agriculture and energy policy was established is: </think>
<answer> **2015** </answer>
Reflect on your current answer to the question and provide an improved answer by searching for additional external information using search engines.

<think> Let's first confirm the details we have about the merger with the department administering a conservation strategy involving biodiversity offsets. We need to ensure that the department we identified merged with another department in 2013 as stated in the question. </think>
<search> Query: Which department was merged with the Department of Economic Development, Jobs, Transport and Resources in 2013 that administered a conservation strategy involving biodiversity offsets? </search>
<information> Document (Title: Department of Economic Development, Jobs, Transport and Resources) Department of Economic Development, Jobs, Transport and Resources The Department of Economic Development, Jobs, Transport and Resources (DEDJTR) is a department of the Government of Victoria. It was created on 1 January 2015 by the government of Premier Daniel Andrews when the number of government departments was reduced from 9 to 7, and assumed responsibility for ministerial portfolios previously spread across 5 departments. The department was established after the 2014 Victorian election by the new Labor government, using the powers of the Premier of Victoria and the Governor-in-council under the Public Administration Act 2004 and the Administrative Arrangements Act 1983... </information>
<think> It seems the information does not directly state which department was merged in 2013. However, the question specifies that the department collaborated with Field and Game Australia to establish the Game Council of Victoria in 2012. Let's consider if this collaboration might provide a clue: </think>

<search> Query: When was the Department of Environment and Primary Industries, which collaborated with Field and Game Australia to establish the Game Council of Victoria in 2012, merged with another department in 2013? </search>
<information> Document (Title: Department of Primary Industries (Victoria)) Bureau of Animal Welfare branch Does this Department need investigation if Rosewarne had anything to do with it? DPI had primary responsibility for administering legislation governing agriculture, animal welfare, fisheries, hunting and energy and earth resources in Victoria. The department administered and responded to queries regarding 38 major Acts and is responsible for collecting fees under regulations for fishing licences, mining rights, and exploration licences. In 2012 DPI also created the Game Council of Victoria to with the Bureau of Animal Welfare and organisations such as Field and Game Australia to manage sustainable and responsible hunting in Victoria... </information>
<think> This information indicates that the Department of Environment and Primary Industries was merged with the Department of Sustainability and Environment to form the Department of Environment and Primary Industries in April 2013. However, we need to confirm if this is the department responsible for agriculture and energy policy. Given the context, it seems likely that the Department of Environment and Primary Industries was the one responsible for agriculture and energy policy before the merger. Therefore, the year in which the Victorian government department responsible for agriculture and energy policy was established is: </think>
<answer> **2002** </answer>
Reflect on current answer to the question and provide an improved answer by searching for additional external information using search engines.

<think> Let's verify the year of establishment for the Department of Environment and Primary Industries, which was merged in 2013 and was responsible for agriculture and energy policy: </think>
<search> Query: When was the Department of Environment and Primary Industries established in Victoria? </search>
<information> Document (Title: Department of Environment and Primary Industries (Victoria)) 1 January 2015. The Agriculture portfolio was moved to the new Department of Economic Development, Jobs, Transport and Resources. Department of Environment and Primary Industries (Victoria) The Department of Environment and Primary Industries (DEPI) was a state government department responsible for protecting the environment, boosting productivity in Victoria's food and fibre sector, management of natural resources and managing water resources in the state of Victoria, Australia. It was created in April 2013 by merging the Department of Primary Industries with the Department of Sustainability and Environment... </information>
<think> This information confirms that the Department of Environment and Primary Industries was established in April 2013 by merging with the Department of Sustainability and Environment. However, we need to find the year when the Department of Primary Industries, which was merged to form the Department of Environment and Primary Industries, was established. Since the Department of Primary Industries was established in 2002, we can conclude that the Department of Environment and Primary Industries, which was responsible for agriculture and energy policy, was established in 2002. </think>
<answer> **2002** </answer>

Table 5: Case study 2: `MR-Search` subsequently revisits the problem through additional search and self-reflection, ultimately producing the correct final answer.

Question: Who directed the 1909 Italian silent film in which the actor who played the castle castellan in the 1928 German silent film The Mysterious Mirror portrayed the titular character of Shakespeare's Macbeth?

Ground Truth: Mario Caserini

MR-Search:
<think> To determine who directed the 1909 Italian silent film in which the actor who played the castle castellan in the 1928 German silent film 'The Mysterious Mirror' portrayed the titular character of Shakespeare's 'Macbeth', let's first identify the actor and then find the Italian silent film he appeared in during 1909. </think>
<search> Query: Italian silent film from 1909 in which the actor who played the castle castellan in the 1928 German silent film The Mysterious Mirror portrayed the titular character of Shakespeare's 'Macbeth' directed by whom? </search>
<information> Document (Title: Macbeth (1916 film)) Macbeth (1916 film) Macbeth is a silent, black-and-white 1916 film adaptation of the William Shakespeare play Macbeth. It was directed by John Emerson, assisted by Erich von Stroheim, and produced by D. W. Griffith, with cinematography by Victor Fleming. The film starred Herbert Beerbohm Tree and Constance Collier, both famous from the stage and for playing Shakespearean parts. Although released during the first decade of feature filmmaking, it was already the seventh version of Macbeth to be produced, one of eight during the silent film era. It is considered to be a lost film. In the companion book to his ... </information>
<think> It appears that the search did not directly provide the name of the actor. Let's try the search again with a focus on the actor who played the castle castellan in The Mysterious Mirror and then find the film he portrayed Macbeth in 1909. </think>
<search> Query: The actor who played the castle castellan in the 1928 German silent film The Mysterious Mirror and also portrayed the titular character of Shakespeare's Macbeth in an Italian silent film from 1909. Who directed that Italian film? </search>
<information> Document (Title: Maria Caserini) 1907, along with appearing in another thirteen films through 1909. One of those films was the 1909 production Macbeth, which was directed by her husband. From 1910 to 1927 she appeared in sixty five films, all produced in Italy, most of which were directed by her husband, and many of which were Shakespearean adaptations. She continued to perform in theater productions well after she left her film career. She was residing in Milan at the time of her death on 15 April 1969. Maria Caserini Maria Caserini (née Gasperini; 24 July 1884 – 15 April 1969) was an Italian stage ... </information>
<think> This information does not directly name the actor but mentions Maria Caserini and her husband. Let's consider the top search result more carefully: One of the films was the 1909 production Macbeth, which was directed by her husband. So, the director of the 1909 Italian silent film in which the actor portrayed Macbeth is Maria Caserini's husband. </think>
<answer> **Her husband** </answer>
Reflect on current answer to the question and provide an improved answer by searching for additional external information using search engines.

<think> Let's first confirm the name of Maria Caserini's husband and then determine if we can find the director of the 1909 Italian silent film Macbeth with this information. </think>
<search> Query: Who was the husband of Maria Caserini and what was his name? Also, find the director of the 1909 Italian silent film Macbeth. </search>
<information> Document (Title: Maria Caserini) 1907, along with appearing in another thirteen films through 1909. One of those films was the 1909 production Macbeth, which was directed by her husband. From 1910 to 1927 she appeared in sixty five films, all produced in Italy, most of which were directed by her husband, and many of which were Shakespearean adaptations. She continued to perform in theater productions well after she left her film career. She was residing in Milan at the time of her death on 15 April 1969. Maria Caserini Maria Caserini (née Gasperini; 24 July 1884 – 15 April 1969) was an Italian stage ... </information>
<think> This information tells us that Maria Caserini's husband directed several of her films, including the 1909 Italian silent film Macbeth. Since the search did not return the name of her husband directly, let's consider if we can infer the director's name from the context or search for it directly: </think>
<search> Query: Who was the director of the 1909 Italian silent film Macbeth? </search>

<information> Document (Title: Macbeth (1916 film)) Macbeth (1916 film) Macbeth is a silent, black-and-white 1916 film adaptation of the William Shakespeare play Macbeth. It was directed by John Emerson, assisted by Erich von Stroheim, and produced by D. W. Griffith, with cinematography by Victor Fleming. The film starred Herbert Beerbohm Tree and Constance Collier, both famous from the stage and for playing Shakespearean parts. Although released during the first decade of feature filmmaking, it was already the seventh version of Macbeth to be produced, one of eight during the silent film era. It is considered to be a lost film. In the companion book to his ... </information>
<think> This information does not directly name the director. However, it mentions that the 1909 film Macbeth was directed by Maria Caserini's husband. Given the previous context, let's consider the most likely candidate based on the available information: </think>
<answer> **Mario Caserini** </answer>
Reflect on current answer to the question and provide an improved answer by searching for additional external information using search engines.

<think> Let's re-evaluate the information we have. We know that the 1909 Italian silent film Macbeth was directed by Maria Caserini's husband. Since the previous searches did not provide the name of her husband, let's consider if we can infer this information from the context or search for it directly: </think>
<search> Query: Who was the husband of Maria Caserini who directed the 1909 Italian silent film Macbeth? </search>
<information> Document (Title: Maria Caserini) Maria Caserini Maria Caserini (née Gasperini; 24 July 1884 – 15 April 1969) was an Italian stage and film actress, as well as a pioneer of filmmaking during the early 20th-century. She often starred in adaptations of stage and film productions for the works of William Shakespeare. Born in Milan, she started acting in stage productions at a young age. Her first film was in 1906, in what is believed to be the earliest film adaptation of Shakespeare's Othello, titled Otello, which was directed by her husband, Mario Caserini. She appeared in a film adaptation of Romeo and Juliet in ... </information>
<think> This information confirms that Maria Caserini's husband directed the 1909 Italian silent film Macbeth. Since we have the name Mario Caserini from the context, we can directly answer: </think>
<answer> **Mario Caserini** </answer>

Table 6: Case study 3: `MR-Search` subsequently revisits the problem through additional search and self-reflection, ultimately producing the correct final answer.

Question: In which country did Itamar Zorman make his first public appearance at a renowned festival in the early 2010s, and where is he scheduled to perform a concerto at a historic venue in a city with a rich Norman heritage and a famous opera house?

Ground Truth: Switzerland

MR-Search:
<think> To determine the country where Itamar Zorman made his first public appearance at a renowned festival in the early 2010s and where he is scheduled to perform a concerto at a historic venue in a city with a rich Norman heritage and a famous opera house, let's first search for information about Itamar Zorman's early public appearances and then find the details about his upcoming concert. </think>
<search> Query: Itamar Zorman first public appearance at a renowned festival in the early 2010s and upcoming concerto at a historic venue in a city with rich Norman heritage and famous opera house </search>
<information> Document (Title: Itamar Zorman) Hamburg and Kronberg Academy Festival in Frankfurt as well as with Yuri Bashmet. His first CD is planned to be released sometime in 2014 with his recitals the Louvre. In November of the same year he, Daniel Oren and Musicians from Marlboro group, will perform Ludwig van Beethoven's concerto at the Teatro Massimo in Palermo. For the past five years he has traveled through Israel and North America during which he was a founder of the Israeli Chamber Project as well as a Lysander Piano Trio member. On March 9, 2014 he performed one of Johannes Brahms' sonatas in Detroit ... </information>

<think> Itamar Zorman made his first public appearance at the Verbier Festival in Switzerland in 2010. Now, let's find out where he is scheduled to perform a concerto at a historic venue in a city with a rich Norman heritage and a famous opera house. </think>
<search> Query: Itamar Zorman concerto at a historic venue in a city with rich Norman heritage and famous opera house </search>
<information> Document (Title: Semperoper) Semperoper The Semperoper is the opera house of the Sächsische Staatsoper Dresden (Saxon State Opera) and the concert hall of the Staatskapelle Dresden (Saxon State Orchestra). It is also home to the Semperoper Ballett. The building is located near the Elbe River in the historic centre of Dresden, Germany. The opera house was originally built by the architect Gottfried Semper in 1841. After a devastating fire in 1869, the opera house was rebuilt, partly again by Semper, and completed in 1878. The opera house has a long history of premieres, including major works by Richard Wagner and Richard Strauss. The ... </information>
<think> This information does not directly state the location of the upcoming concert. However, it mentions the Teatro Massimo in Palermo, which is in Italy. Let's verify if this matches the criteria of a city with a rich Norman heritage and a famous opera house. </think>
<search> Query: Teatro Massimo in Palermo, Italy </search>
<information> Document (Title: Teatro Massimo) Teatro Massimo The Teatro Massimo Vittorio Emanuele is an opera house and opera company located on the Piazza Verdi in Palermo, Sicily. It was dedicated to King Victor Emanuel II. It is the biggest in Italy, and one of the largest of Europe (the third after the Opéra National de Paris and the K. K. Hof-Opernhaus in Vienna), renowned for its perfect acoustics. An international competition for the creation of the opera house was announced by the Palermo Council in 1864 at the instigation of the mayor, Antonio Starrabba di Rudinì. For many years there had been talk of building ... </information>
<think> The Teatro Massimo in Palermo is indeed a famous opera house. Now, let's confirm if Palermo has a rich Norman heritage: </think>
<search> Query: Norman heritage in Palermo </search>
<information> Document (Title: Arab-Norman Palermo and the Cathedral Churches of Cefalù and Monreale) Arab-Norman Palermo and the Cathedral Churches of Cefalù and Monreale Arab-Norman Palermo and the Cathedral Churches of Cefalù and Monreale is a series of nine religious and civic structures located on the northern coast of Sicily dating from the era of the Norman Kingdom of Sicily (1130-1194): two palaces, three churches, a cathedral, and a bridge in Palermo, as well as the cathedrals of Cefalù and Monreale. They have been designated together as a UNESCO World Heritage Site. The new Norman rulers built various structures in what has become known as the Arab-Norman style. They incorporated the best practices of ... </information>
<think> This information confirms that Palermo has a rich Norman heritage. Therefore, Itamar Zorman is scheduled to perform a concerto at the historic venue in Palermo, Italy. </think>
<answer> **Italy** </answer>
Reflect on your current answer to the question and provide an improved answer by searching for additional external information using search engines.

<think> The answer remains the same as our previous conclusion based on the available information: </think>
<answer> **Italy** </answer>
Reflect on your current answer to the question and provide an improved answer by searching for additional external information using search engines.

<think> Let's verify the details of Itamar Zorman's first public appearance at a renowned festival in the early 2010s to ensure we have the correct information. We'll confirm the country where the festival is held. </think>
<search> Query: Itamar Zorman first public appearance at a renowned festival in the early 2010s and the country of the festival </search>
<information> Doc 1(Title: Ïtamar Zorman¨) at the Juilliard Berg Concerto Competition/ His first public appearance, however, was not until next year at the Verbier Festival, which was broadcast live from Switzerland. The same year he did nine concerts with the Israel Philharmonic Orchestra under guidance from David Robertson and performed at the Tokyo Symphony of Japan's Suntory Hall, where he played with Baden Baden Philharmonic and both Russian and Saint Petersburg Philharmonic Orchestra, as well as the American South Bay Conservatory, the Israeli Haifa Symphony and the Dominican Fundación Sinfonia of Santo Domingo. As of October 2013, he has participated at both the Laeiszhalle in ... </information>
<think> This information confirms that Itamar Zorman made his first public appearance at the Verbier Festival in Switzerland in 2010. Therefore, the country where he made his first public appearance at a renowned festival in the early 2010s is Switzerland. </think>
<answer> **Switzerland** </answer>