SCALING SEARCH-AUGMENTED LLM REASONING VIA ADAPTIVE INFORMATION CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Search-augmented reasoning agents interleave multi-step reasoning with external information retrieval, but uncontrolled retrieval often leads to redundant evidence, context saturation, and unstable learning. Existing approaches typically rely on outcome-based reinforcement learning (RL), where sparse, delayed rewards provide limited guidance for regulating when, how much, and at what granularity information should be acquired. We propose DEEPCONTROL, a framework for adaptive information control grounded in a formal notion of *information utility*, which quantifies the state-dependent marginal value of retrieved evidence for ongoing reasoning. Building on this utility, we introduce retrieval continuation and granularity control mechanisms that selectively decide whether retrieval should proceed and which parts of hierarchical information to expand. An annealed control strategy further enables the agent to internalize effective information acquisition behaviors during training. Extensive experiments across seven benchmarks demonstrate that our method consistently outperforms strong outcome-based RL baselines and retrieval-free or retrieval-based reasoning methods without explicit information control. In particular, compared with Search-R1, a strong outcome-based RL baseline, our approach improves average performance by +9.4 and +8.6 points on Qwen2.5-7B and Qwen2.5-3B, respectively. Beyond performance, our analysis reveals how information utility evolves with retrieval depth and training scale, shedding light on efficiency–performance trade-offs in large-scale post-training for search-augmented reasoning agents.

1 INTRODUCTION

Recent advances have enabled deep research agents that interleave multi-step reasoning with external information acquisition, allowing language models to solve complex, knowledge-intensive tasks beyond their parametric knowledge (Zheng et al., 2025; Du et al., 2025; Huang et al., 2025). As these agents are deployed in increasingly large information environments, where the amount, length, and structural complexity of retrievable content grow substantially, their performance is no longer limited by search availability or reasoning capacity alone. Instead, a new bottleneck emerges: uncontrolled information acquisition. In practice, repeatedly retrieving more evidence often leads to context saturation, redundant or noisy information accumulation, and interference between reasoning and retrieved content, ultimately degrading decision quality rather than improving it (Yu et al., 2024; Jin et al., 2025). Crucially, these failures indicate that more retrieval does not necessarily lead to better reasoning.

To mitigate these issues, prior work (Jin et al., 2025; Zheng et al., 2025) has predominantly relied on outcome-based reinforcement learning (Schulman et al., 2017; Guo et al., 2025), using final answer correctness as the sole training signal to guide both reasoning and retrieval decisions. However, such outcome-only learning signals introduce fundamental limitations when regulating information acquisition. In particular, agents often exhibit suboptimal retrieval behaviors: they may over-retrieve when evidence is weak or queries are poorly specified, accumulating unnecessarily long contexts instead of relying on internal knowledge; conversely, they may terminate retrieval prematurely even when additional evidence remains beneficial. These issues are exacerbated in long-horizon reasoning settings, where sparse outcome rewards provide limited guidance for intermediate retrieval decisions, leading to unstable training and inefficient exploration (Xiong et al., 2025a). Importantly, these failures do not arise from inadequate search capability, but from the inability of outcome-only learning signals to regulate when, how much, and at what granularity information should be acquired. What is missing, therefore, is explicit and adaptive control over information acquisition. The utility of retrieved information is inherently state-dependent and evolves over the course of reasoning. Effective agents must reason at multiple levels of granularity, selectively expanding fine-grained details only when they are expected to provide marginal benefit. Retrieval should thus be incremental, selective, and interruptible, allowing the agent to balance the benefits of additional information against its computational and contextual costs. Without such control, scaling search-augmented reasoning agents to large information spaces remains fundamentally brittle.

In this work, we introduce a framework for adaptive information control in search-augmented reasoning agents. Rather than relying solely on outcome-based reinforcement learning, our approach equips the agent with explicit control signals that guide information acquisition during reasoning. The agent is trained to regulate retrieval granularity, decide when to expand additional evidence, and determine when to halt further retrieval. These control mechanisms operate alongside standard online RL optimization, allowing the agent to retain the flexibility of learning from interaction while correcting systematic retrieval failures. Our framework complements outcome-based reinforcement learning rather than replacing it. We continue to optimize the agent using standard online RL objectives, but augment training with structured control signals that intervene when retrieval behavior is misaligned with information utility. This design leads to more efficient exploration, reduced context waste, and improved training stability, particularly in long-horizon reasoning scenarios. As a result, the agent learns not only how to search, but also how to control the flow of external information during reasoning.

In summary, our main contributions are threefold:

- We propose a formal definition of *information utility* for search-augmented reasoning, which characterizes the marginal value of retrieved information under a given reasoning state. The utility captures two complementary aspects, novelty and effectiveness, and is empirically shown to distinguish useful evidence from irrelevant or redundant retrievals, providing a principled basis for information acquisition control.
- Building on information utility, we introduce two information control mechanisms: *retrieval continuation control* and *granularity control*. The former adaptively determines whether retrieval should continue or terminate, avoiding both premature stopping and over-retrieval, while the latter enables selective expansion of high-utility content within hierarchical information structures. We further adopt an *annealed control strategy* that gradually removes external control during training, allowing the model to internalize effective information acquisition behaviors.
- We conduct extensive experiments across multiple tasks, datasets, and model scales, demonstrating that our approach consistently outperforms existing search-augmented reasoning methods in reasoning accuracy, training stability, and computational efficiency across diverse information scales and reasoning complexities.

Together, these results underscore the importance of adaptive information control for scaling search-augmented reasoning agents to complex, real-world information environments.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

We consider a search-augmented reasoning agent that solves complex queries by interleaving multi-step reasoning with external information retrieval. Given a task $u \sim \mathbb{P}(\mathcal{U})$, the agent governed by a policy π_θ interacts with a search engine $\mathcal{R}$ and maintains a reasoning state s_t, i.e., the accumulated context, including retrieved evidence and intermediate reasoning. Specifically, at each time step t, the policy samples a structured action $a_t = (h_t, \alpha_t, \xi_t)$ according to $a_t \sim \pi_\theta(\cdot \mid u, s_t)$, where: (i) h_t denotes natural-language reasoning tokens, (ii) α_t denotes the action-type tokens (e.g., `retrieve`), and (iii) ξ_t represents action parameters, such as the search query issued to the search engine $\mathcal{R}$. A rollout trajectory is a sequence of states and actions: $\tau = (s_0, a_0, \ldots, a_{T-1}, s_T)$. The episode terminates when the agent outputs a final answer or when the maximum number of steps is reached.

2.2 ONLINE RL WITH SEARCH-AUGMENTED REASONING AGENTS

Online RL alternates between a *rollout phase*, in which trajectories are generated with the current policy, and an *update phase*, in which the policy is optimized using collected rollouts. We optimize the agent to maximize task success while regularizing deviation from a reference policy π_{ref}.

Proximal Policy Optimization. Proximal Policy Optimization (PPO) (Schulman et al., 2017) is a widely used actor–critic algorithm for LLM post-training (Ouyang et al., 2022). For LM-based agents, PPO optimizes the policy by maximizing the following objective:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{u \sim \mathbb{P}(\mathcal{U}),\, \tau \sim \pi_{\theta_{\text{old}}}} \left[\min\left(\frac{\pi_\theta(a_t \mid u, s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid u, s_t)} A_t,\ \text{clip}\left(\frac{\pi_\theta(a_t \mid u, s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid u, s_t)}, 1 - \epsilon, 1 + \epsilon \right) A_t \right) \right], \quad (1)$$

where π_θ and $\pi_{\theta_{\text{old}}}$ denote the current and previous policy models, respectively. The hyperparameter ϵ controls the clipping range and stabilizes training. The advantage estimate A_t is computed using Generalized Advantage Estimation (GAE) (Schulman et al., 2015), based on future rewards $\{r_{\geq t}\}$ and a learned value function V_ζ.

Adaptations for search-augmented reasoning. In search-augmented reasoning, retrieved content is produced by an external search engine rather than the policy itself. As a result, policy-gradient updates apply only to tokens generated by the language model. Existing approaches use a final outcome-based reward $\mathbb{I}[y_{\text{pred}} = y_{\text{gold}}]$, which evaluates whether the agent's final prediction y_{pred} exactly matches the gold answer y_{gold}, typically using Exact Match (EM).

Discussion on the weakness of outcome-based RL training. The above approach of search-augmented reasoning with outcome-based RL training enables the agent to learn how to use search tools, but introduces several issues:

1) **Suboptimal Search Behavior.** The search behavior of agents is often suboptimal. For example, when relevant evidence is unavailable or the query is poorly specified, the agent may over-retrieve, accumulating unnecessarily long contexts, instead of answering based on existing information and internal knowledge. While outcome-based reinforcement learning can partially mitigate this issue, learning remains inefficient in the absence of explicit control signals.
2) **Information Overload.** Most existing approaches (Lin et al., 2023; Yu et al., 2024; Jin et al., 2025) naively append raw retrieved content to the context, which can quickly overwhelm the context window, especially when sources are long (e.g., webpages or academic papers). To alleviate this, they often adopt a small top-k (e.g., $k = 3$), which risks missing critical evidence even when the retrieval query is correct, or increase the maximum context length (e.g., 32K tokens), substantially raising training and inference costs. As a result, these limitations significantly hinder the applicability of such methods to complex real-world scenarios.
3) **Unstable Training.** Outcome-based RL provides sparse supervision, making policy optimization highly sensitive to individual mistakes along long reasoning trajectories. This challenge is exacerbated when initializing from weak base models, where inaccurate exploration further destabilizes training.

3 ADAPTIVE INFORMATION CONTROL

3.1 INFORMATION UTILITY

The value of external information acquisition is inherently *state-dependent* and must be assessed relative to the agent's current reasoning state. In our framework, information acquisition is organized into discrete *search steps* (Figure 1). Each search step starts with a retrieval action and is followed by a variable number of expansion actions that selectively refine the retrieved information as needed (Section 3.2). We treat each search step as a single unit for utility estimation, abstracting away intermediate expansion states.

We distinguish between two levels of indexing. Let $t = 0, 1, \ldots, T-1$ index primitive actions (e.g., `retrieve`, `expand`, `answer`), and let $l = 0, 1, \ldots, L-1$ index search steps. Let t_l denote the primitive step at which the l-th retrieval is executed. The l-th search step starts at t_l and includes the

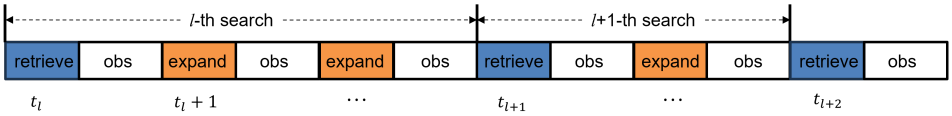


Figure 1: Definition of a search step. A search step starts with a retrieval action and includes all subsequent expansion actions until the next retrieval or termination.

retrieval action together with all subsequent expansion actions until the next retrieval or termination. Let t_{l+1} denote the primitive step of the next retrieval (or the termination boundary), so that all expansions triggered by the l-th retrieval are completed by step $t_{l+1} - 1$.

Let u denote the task, and let s_{t_l} denote the agent's reasoning state immediately before executing the l-th retrieval. We denote by e_l the retrieval output at search step l. We define the information utility of the l-th search step as

$$U(e_l \mid u, s_{t_l}) = \rho \cdot \mathrm{Novelty}(e_l \mid s_{t_l}) + (1-\rho) \cdot \mathrm{Effectiveness}(e_l \mid u, s_{t_l}). \tag{2}$$

where $\rho \in [0, 1]$ balances the contribution of novelty and effectiveness. For notational simplicity, we write $U(e_l)$ instead of $U(e_l \mid u, s_{t_l})$ in subsequent sections.

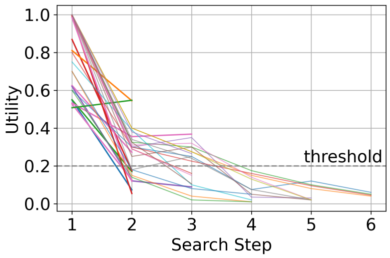


Figure 2: Information utility varies across search steps and does not monotonically increase with additional retrieval, motivating explicit continuation control.

The proposed information utility satisfies several desirable properties for regulating retrieval, including state-dependence and diminishing marginal gains. Unlike outcome-based rewards that only provide sparse terminal signals, information utility offers dense, state-dependent feedback for intermediate retrieval decisions. Detailed definitions, theoretical properties, and additional discussion are provided in Section A.

3.2 GRANULARITY CONTROL VIA HIERARCHICAL SELECTIVE EXPANSION

In real-world settings, retrieved information can be voluminous, making it impractical to inject all content into the agent context. Moreover, fine-grained details are not uniformly useful across reasoning stages. We therefore introduce *granularity control*, which exposes retrieved information at a coarse level and selectively refines finer-grained content only when beneficial.

Under granularity control, **retrieval and information refinement are decoupled**. The agent first retrieves coarse-grained information and then performs explicit `expand` actions to access more detailed evidence as needed. Retrieved information is organized hierarchically, enabling selective

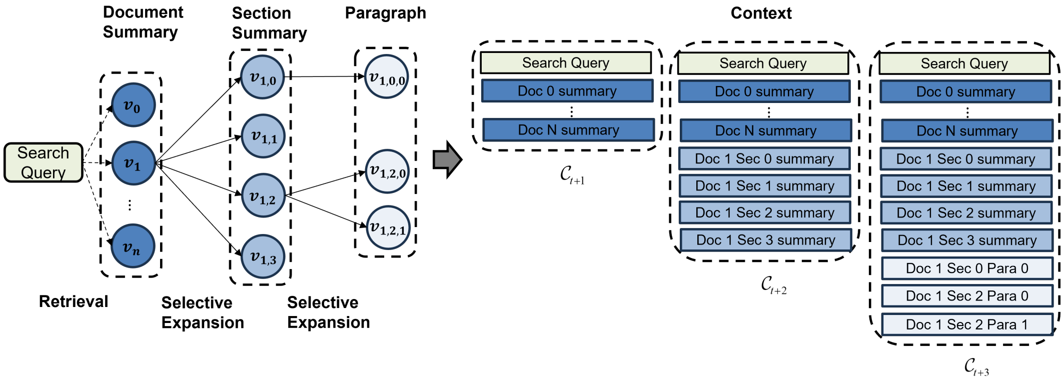


Figure 3: Hierarchical granularity control via selective expansion, where the agent incrementally refines retrieved information from coarse summaries to finer-grained content as needed.

expansion from high-level summaries to finer-grained units and focusing the context budget on information most relevant to the current reasoning state.

As illustrated in Figure 3, the agent initializes its context with coarse summaries and incrementally refines them through a small number of targeted expansion steps, reducing unnecessary context growth while preserving access to detailed evidence when required. During training, expansion decisions are guided by information utility, which provides state-dependent supervision over which parts of the hierarchy to refine. Formal definitions are deferred to Section C.1.

3.3 SEARCH CONTINUATION CONTROL

By default, the agent autonomously decides whether to continue searching based on its internal reasoning state, which is often suboptimal: it may terminate search prematurely or overcommit to unnecessary retrieval. We therefore model *search continuation* as an explicit control decision, using information utility as a monitoring signal to correct systematic misjudgments (Figure 2).

We introduce two complementary interventions. *Termination control* halts search when the utility of recent retrievals consistently remains low, preventing over-retrieval. Conversely, *continuation control* forces additional retrieval when recent utility remains high but the agent exhibits low confidence in its current answer. Importantly, information utility does not replace the agent's policy, but acts as a lightweight supervisory signal that triggers corrective control only when necessary, avoiding excessive intervention. Formal definitions are deferred to Section C.2.

3.4 REINFORCEMENT LEARNING WITH INFORMATION CONTROL

External control signals can stabilize early-stage reinforcement learning by correcting systematic errors in information acquisition, but must be internalized to improve intrinsic capabilities at test time. We address this with an annealed *control-forcing* reinforcement learning scheme that transitions from guided to fully autonomous behavior.

During training, the agent alternates between two rollout modes. In the *controlled* mode, a lightweight controller monitors information utility and injects explicit control signals when abnormal retrieval behavior is detected, which the policy conditions on when generating actions. In the *uncontrolled* mode, the policy operates autonomously without external intervention. Figure 4 shows example trajectories under the two modes.

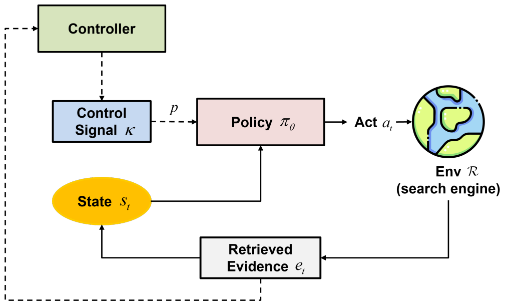


Figure 4: Trajectories generated in rollout mode with and without information control.

To prevent over-reliance on control, we progressively reduce the frequency of controlled rollouts over training, ensuring that the final policy performs reliably in the autonomous setting. We optimize the policy using a composite reward that combines outcome correctness with lightweight regularization on tool usage and retrieval behavior. Formal definitions are deferred to Section C.3.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets We evaluate DEEPCONTROL on seven benchmarks: **General QA** (NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2022)) and **Multi-hop QA** (HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), Musique (Trivedi et al., 2022b), Bamboogle (Press et al., 2022)).

Baselines We mainly compare DEEPCONTROL against three groups of baselines: **(i) Inference without Retrieval**: Direct inference and CoT (Wei et al., 2022); **(ii) Inference with Retrieval**:

Table 1: Main results with best performance in bold. $^\dagger/^\star$ represents in-domain/out-domain datasets.

Methods	**General QA**			**Multi-Hop QA**				
	NQ†	**TriviaQA***	**PopQA***	**HotpotQA**†	**2wiki***	**Musique***	**Bamboogle***	**Avg.**
Qwen2.5-7b-Base/Instruct								
Direct Inference	0.134	0.408	0.140	0.183	0.250	0.031	0.120	0.181
CoT	0.048	0.185	0.054	0.092	0.111	0.022	0.232	0.106
IRCoT	0.224	0.478	0.301	0.133	0.149	0.072	0.224	0.239
Search-o1	0.151	0.443	0.131	0.187	0.176	0.058	0.296	0.206
RAG	0.349	0.585	0.392	0.299	0.235	0.058	0.208	0.304
SFT	0.318	0.354	0.121	0.217	0.259	0.066	0.112	0.207
R1-base	0.297	0.539	0.202	0.242	0.273	0.083	0.296	0.276
R1-instruct	0.270	0.537	0.199	0.237	0.292	0.072	0.293	0.271
Rejection Sampling	0.360	0.592	0.380	0.331	0.296	0.123	0.355	0.348
Search-R1-base	0.480	0.638	0.457	0.433	0.382	0.196	0.432	0.431
Search-R1-instruct	0.393	0.610	0.397	0.370	0.414	0.146	0.368	0.385
Ours	**0.558**	**0.682**	**0.521**	**0.471**	**0.439**	**0.221**	**0.458**	**0.479**
Qwen2.5-3b-Base/Instruct								
Direct Inference	0.106	0.288	0.108	0.149	0.244	0.020	0.024	0.134
CoT	0.023	0.032	0.005	0.021	0.021	0.002	0.000	0.015
IRCoT	0.111	0.312	0.200	0.164	0.171	0.067	0.240	0.181
Search-o1	0.238	0.472	0.262	0.221	0.218	0.054	**0.320**	0.255
RAG	0.348	0.544	0.387	0.255	0.226	0.047	0.080	0.270
SFT	0.249	0.292	0.104	0.186	0.248	0.044	0.112	0.176
R1-base	0.226	0.455	0.173	0.201	0.268	0.055	0.224	0.229
R1-instruct	0.210	0.449	0.171	0.208	0.275	0.060	0.192	0.224
Rejection Sampling	0.294	0.488	0.332	0.240	0.233	0.059	0.210	0.265
Search-R1-base	0.406	0.587	0.435	0.284	0.273	0.049	0.088	0.303
Search-R1-instruct	0.341	0.545	0.378	0.324	0.319	0.103	0.264	0.325
Ours	**0.533**	**0.645**	**0.512**	**0.402**	**0.371**	**0.118**	0.298	**0.411**

RAG (Lewis et al., 2020), IRCoT (Trivedi et al., 2022a), and Search-o1 (Li et al., 2025); **(iii) Fine-Tuning-Based Methods**: SFT (Chung et al., 2024), RL without search (R1) (Guo et al., 2025), rejection sampling with search (Ahn et al., 2024) and Search-R1 (Jin et al., 2025). For R1, rejection sampling and Search-R1, we use the fine-tuned version from (Jin et al., 2025). Across all methods, we use the same retriever, corpus, effective retrieval budget, training data, and pretrained LLMs.

Implementation details We conduct experiments with Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct (Yang et al., 2024). For retrieval, we use the 2018 Wikipedia dump (Karpukhin et al., 2020) with E5 (Wang et al., 2022) as the retriever. Unlike prior methods that append raw retrieved passages to the context, our approach employs hierarchical selective expansion. For fair comparison with existing retrieval-based baselines (Lin et al., 2023), we control the effective evidence budget across methods.

For training, following (Jin et al., 2025), we merge the training sets of NQ and HotpotQA into a unified dataset and optimize the agent using reinforcement learning. Evaluation is conducted on the test or validation sets of seven benchmarks to assess both in-domain and out-of-domain performance, using Exact Match (EM) as the metric following (Yu et al., 2024). Additional training details, hyperparameters, and ablations are provided in Section D.

4.2 MAIN RESULTS

The main results comparing DEEPCONTROL with baseline methods across seven datasets are summarized in Table 1, with qualitative examples provided in Section E. We draw the following key observations. **(1) DEEPCONTROL consistently outperforms strong baselines.** Compared with Search-R1, DEEPCONTROL achieves average improvements of +9.4 and +8.6 points on Qwen2.5-7B and Qwen2.5-3B, respectively, and consistently improves performance across both in-distribution and out-of-distribution benchmarks. **(2) Explicit information control is critical for retrieval-based reasoning.** DEEPCONTROL outperforms RL-based reasoning both without retrieval (R1) and with retrieval but without information control (Search-R1), demonstrating that ef-

Table 2: Ablation study results. We evaluate the impact of different control signals.

Method	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	Avg.
Qwen2.5-3b-Instruct								
DEEPCONTROL	**0.533**	**0.645**	**0.512**	**0.402**	**0.371**	**0.118**	**0.298**	**0.411**
w/o Granularity Control	0.470	0.580	0.440	0.340	0.310	0.080	0.230	0.364
w/o Search Continuation Control	0.490	0.610	0.460	0.360	0.340	0.100	0.260	0.380
w/o Both	0.406	0.545	0.378	0.284	0.273	0.049	0.088	0.303

fective reasoning requires not only access to external information but also explicit control over how it is used during search. **(3) Larger models benefit more from search learning.** The 7B variant exhibits a larger performance margin over Search-R1 than the 3B variant, indicating that larger models are more effective at learning and exploiting search-based reasoning strategies.

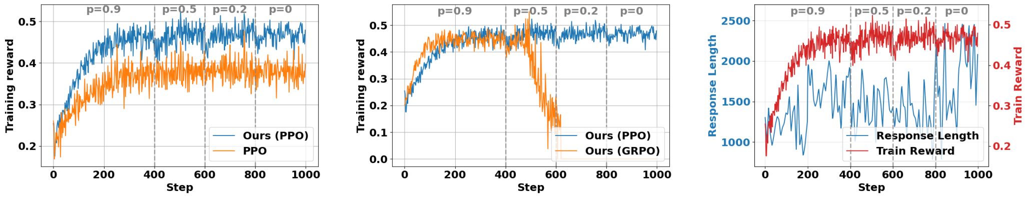
4.3 ANALYSIS

Effect of Information Control on Online RL Training. We compare DEEPCONTROL with vanilla PPO under identical data, reward, and hyperparameter settings. As shown in Figure 5(a) and Table 3, DEEPCONTROL consistently outperforms vanilla PPO. Information control provides corrective guidance during early training, preventing suboptimal retrieval behaviors when the policy is immature, and is gradually internalized as training progresses. Overall, DEEPCONTROL improves performance by 8.3% on average, demonstrating that information control substantially enhances learning efficiency in online RL.

RL Algorithms under Annealed Control. We evaluate PPO and GRPO under the same annealed control-forcing setting. Figure 5(b) and Table 4 show that while GRPO converges faster in early training, it suffers from reward collapse under control annealing. In contrast, PPO remains stable throughout training and achieves higher final performance without control signals, indicating greater robustness in this setting.

Response Length Dynamics. Figure 5(c) shows the evolution of response length during training. Early training exhibits increased actions due to guided search and expansion. As training stabilizes and control is removed, both response length and performance converge, indicating successful internalization of controlled search behaviors.

Ablation Study. Table 2 presents ablations on Qwen2.5-3B-Instruct across seven benchmarks. Removing either granularity control or search continuation control consistently degrades performance, while removing both leads to substantial drops across all datasets. Granularity control is particularly important for multi-hop and long-context tasks, where resolution-adaptive refinement limits context growth, whereas continuation control mainly benefits datasets with heterogeneous evidence quality by preventing premature termination and over-retrieval. These results highlight the complementary roles of the two components in enabling efficient information acquisition.



(a) DEEPCONTROL vs. PPO (b) PPO vs. GRPO (c) Response length

Figure 5: (a) DEEPCONTROL vs. PPO: our approach reaches higher reward throughout training under the same optimization setup. (b) PPO vs. GRPO: GRPO leads to reward collapse, while PPO shows steadier optimization and maintains stable performance. (c) Response-length behavior during training: the average response length and training reward evolve together over time, with response length increasing at early stages and tending to stabilize later.

5 RELATED WORK

Large Language Models with Retrieval Large language models (LLMs) have demonstrated strong reasoning and coding capabilities, yet often suffer from limited factual coverage and hallucinations (Zhang et al., 2023). To address these limitations, external retrieval systems are commonly integrated to provide additional information. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) incorporates retrieved documents directly into the model context, while tool-based approaches treat search engines as external tools invoked during reasoning. Representative methods include IRCoT (Trivedi et al., 2022a), ReAct (Yao et al., 2023), Toolformer (Schick et al., 2023) and Search-R1 (Jin et al., 2025), which enable LLMs to interleave reasoning steps with search engine calls. Despite their success, existing retrieval-augmented approaches largely assume that acquiring more information is beneficial. Retrieved content is typically appended to the context in a fixed or heuristic manner, without explicitly modeling its marginal utility. As a result, these methods are prone to context saturation, redundant evidence accumulation, and noisy reasoning, especially when operating over large information spaces. In contrast, our work focuses on explicitly regulating information acquisition by modeling the utility of retrieved evidence and controlling both the amount and granularity of information exposed to the agent.

Reinforcement Learning for LLM Reasoning and Tool Use Reinforcement learning has been widely adopted to optimize LLMs for complex behaviors, including reasoning and tool use. RLHF (Ouyang et al., 2022) and its variants (Rafailov et al., 2023) use preference-based rewards to align model outputs with human expectations, while recent work has shown that outcome-based RL can enable LLMs to acquire advanced reasoning skills using only task-level rewards (Shao et al., 2024; Guo et al., 2025). Several methods apply RL to tool-augmented settings, allowing agents to learn when and how to invoke external tools. However, most existing RL-based approaches rely on sparse, outcome-level supervision. While such signals are sufficient for learning final-answer correctness, they provide limited guidance for intermediate decisions, such as whether to continue retrieval, how much information to acquire, or when to stop. As a result, agents often exhibit suboptimal retrieval behaviors, including over-retrieval and premature termination. Our work addresses this limitation by augmenting outcome-based RL with explicit information control signals derived from information utility, enabling more effective regulation of retrieval behaviors during training.

Information Control in Search-Augmented Reasoning Effective exploration is a central challenge in reinforcement learning. Prior work has proposed intrinsic rewards, count-based exploration, and curiosity-driven methods to encourage novelty and state coverage (Pathak et al., 2017; Bellemare et al., 2016). In the context of LLMs, recent studies have explored structured exploration strategies and process-level supervision to improve reasoning diversity and stability (Xiong et al., 2025b;a). While these approaches aim to improve exploration efficiency, they do not explicitly address information acquisition in search-augmented reasoning. In particular, existing methods do not model the utility of retrieved information nor provide mechanisms to control retrieval continuation and granularity. Our work bridges this gap by introducing a utility-driven framework that treats information acquisition as a controllable process, and by using annealed control to enable the agent to internalize effective exploration and retrieval behaviors.

6 CONCLUSION

In this paper, we introduce an adaptive information control framework based on information utility for regulating information acquisition in search-augmented reasoning agents. By modeling the marginal value of retrieved information under different reasoning states, the framework enables explicit control over retrieval continuation and information granularity, leading to more effective use of external information. During online reinforcement learning, control signals are combined with an annealed strategy, allowing the model to gradually internalize appropriate information acquisition behaviors without external intervention. Experimental results across multiple tasks and model scales show improvements in reasoning accuracy, training stability, and computational efficiency, highlighting the importance of treating information acquisition as a controllable and learnable process. Looking forward, this work opens several promising directions. Future research may explore richer definitions of information utility, uncertainty-aware retrieval and stopping criteria, and extensions to multi-tool and multimodal reasoning settings.

REFERENCES

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, 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.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 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*, 2025.

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.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

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*, 2025.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*, 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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

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

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*, 2022a.

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, 2022b.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Siheng Xiong, Ali Payani, and Faramarz Fekri. Enhancing long chain-of-thought reasoning through multi-path plan aggregation. *arXiv preprint arXiv:2510.11620*, 2025a.

Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. Deliberate reasoning in language models as structure-aware planning with an accurate world model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31900–31931, 2025b.

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 Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184, 2024.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

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.

A INFORMATION UTILITY

The value of external information acquisition is inherently *state-dependent* and must be assessed relative to the agent's current reasoning state. We formalize this notion through *information utility*, which measures the marginal value of newly acquired information for the downstream task.

As described in Section 3.1, information acquisition is organized at the level of *search steps*. We distinguish between two levels of indexing: let $t = 0, 1, \ldots, T-1$ index primitive actions (e.g., `retrieve`, `expand`, `answer`), and let $l = 0, 1, \ldots, L-1$ index search steps, each corresponding to a single retrieval event. Let t_l denote the primitive step at which the l-th retrieval is executed. The l-th search step starts at t_l and includes the retrieval action together with all subsequent expansion actions until the next retrieval or termination. Let t_{l+1} denote the primitive step of the next retrieval (or the termination boundary), so that all expansions triggered by the l-th retrieval are completed by step $t_{l+1} - 1$.

Let u denote the task, and let s_{t_l} denote the agent's reasoning state immediately before executing the l-th retrieval. We denote by $\mathcal{C}_t$ the *injected node set* in the agent context after primitive step t, which may include both internal nodes and leaf nodes under hierarchical granularity control.

Retrieval output vs. injected evidence. Under granularity control, retrieval exposes a *hierarchical evidence structure*, while expansions determine which nodes are actually injected into the context. We denote by e_l the *retrieval output* at the l-th search step:

$$e_l \triangleq \{\mathcal{G}_l^{(i)}\}_{i=1}^{k}, \qquad \mathcal{G}_l^{(i)} = (\mathcal{V}_l^{(i)}, \mathcal{E}_l^{(i)}), \tag{3}$$

where each retrieved source is a rooted tree with node set $\mathcal{V}_l^{(i)}$ and directed refinement edges $\mathcal{E}_l^{(i)}$.

Expansions triggered by the l-th retrieval inject a subset of nodes from the retrieved hierarchies into the context, causing the injected set $\mathcal{C}_t$ to grow during the interval $t \in [t_l,\, t_{l+1} - 1]$. We quantify the *net injected nodes* contributed by the l-th search step as the set difference

$$\Delta\mathcal{C}_l \triangleq \mathcal{C}_{t_{l+1}-1} \setminus \mathcal{C}_{t_l - 1}. \tag{4}$$

By construction, $\Delta\mathcal{C}_l$ captures the aggregate information injected due to the l-th retrieval and its subsequent expansions, abstracting away intermediate refinement states.

We additionally define the *retrieved leaf pool* for novelty computation as

$$\tilde{\mathcal{L}}_l \triangleq \mathrm{Leaves}(e_l), \qquad \tilde{\mathcal{L}}_{<l} \triangleq \bigcup_{j<l} \tilde{\mathcal{L}}_j, \tag{5}$$

i.e., $\tilde{\mathcal{L}}_l$ contains *all* leaf nodes in the retrieved hierarchies at search step l, regardless of whether they are injected.

Since injected nodes are selected from the retrieved hierarchies, we have $\Delta\mathcal{C}_l \subseteq \bigcup_{i=1}^{k} \mathcal{V}_l^{(i)}$, and the injected leaf nodes are a subset of the retrieved leaf pool: $\mathrm{Leaves}(\Delta\mathcal{C}_l) \subseteq \tilde{\mathcal{L}}_l$.

Information utility. We define the information utility of the l-th search step as

$$U(e_l) = \rho \cdot \mathrm{Novelty}(e_l \mid s_{t_l}) + (1 - \rho) \cdot \mathrm{Effectiveness}(e_l \mid u, s_{t_l}), \tag{6}$$

where $\rho \in [0, 1]$ balances the contribution of novelty and effectiveness. Concretely, we instantiate $\mathrm{Novelty}(e_l \mid s_{t_l}) \triangleq \mathrm{Novelty}(\tilde{\mathcal{L}}_l \mid \tilde{\mathcal{L}}_{<l})$ and $\mathrm{Effectiveness}(e_l \mid u, s_{t_l}) \triangleq \mathrm{Effectiveness}(\Delta\mathcal{C}_l \mid u, s_{t_l})$. This design decouples *coverage* (novelty over the full retrieved leaf pool) from *impact* (effectiveness of what is actually injected), enabling the controller to detect redundant retrieval even when the agent chooses not to expand those leaves.

Novelty. Under hierarchical granularity control, retrieved information is organized as a multi-resolution tree, where internal nodes correspond to coarse representations (e.g., document or section summaries) and leaf nodes correspond to fine-grained evidence units that contain concrete factual content (e.g., paragraphs). We define novelty at the level of leaf nodes, and compute it over the *entire* leaf pool returned by retrieval.

Each leaf node is embedded into a shared semantic space using the E5 encoder (Wang et al., 2022). For each newly retrieved leaf node $v \in \tilde{\mathcal{L}}_l$, we identify its k_{nn} nearest neighbors among leaf nodes retrieved in prior search steps, denoted by $\tilde{\mathcal{L}}_{<l}$, and compute the average cosine similarity

$$\mathrm{sim}(v) = \frac{1}{k_{\mathrm{nn}}} \sum_{v' \in \mathrm{KNN}\,(v,\, \tilde{\mathcal{L}}_{<l},\, k_{\mathrm{nn}})} \cos(v, v'), \tag{7}$$

which estimates the degree to which the content of v overlaps with previously retrieved evidence. We define the novelty of leaf node v as

$$\mathrm{Novelty}(v) = 1 - \mathrm{sim}(v), \tag{8}$$

and aggregate novelty across the search step by averaging over the retrieved leaf pool:

$$\mathrm{Novelty}(\tilde{\mathcal{L}}_l \mid \tilde{\mathcal{L}}_{<l}) = \frac{1}{|\tilde{\mathcal{L}}_l|} \sum_{v \in \tilde{\mathcal{L}}_l} \big(1 - \mathrm{sim}(v)\big). \tag{9}$$

By restricting novelty evaluation to leaf nodes, this formulation measures redundancy at the level of concrete evidence, while avoiding spurious similarity between fine-grained content and coarse summaries.

Effectiveness. While novelty captures whether newly retrieved information introduces previously unseen content, effectiveness measures whether the information injected by expansions is *useful* for the downstream task, i.e., whether it meaningfully changes the model's belief over candidate answers. Unlike novelty, effectiveness is computed with respect to the full set of injected nodes contributed by the search step, $\Delta\mathcal{C}_l$, which may include both internal and leaf nodes.

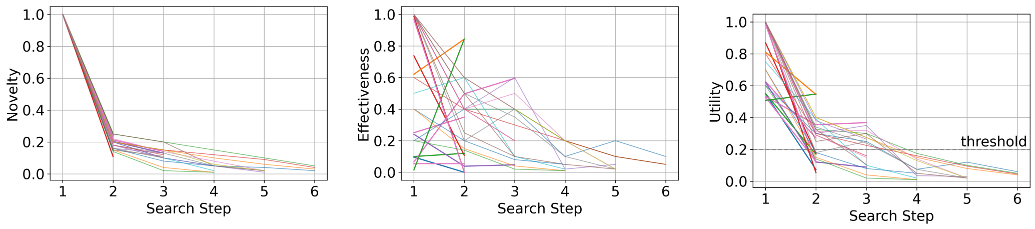
Let $\mathcal{Y}$ denote a finite set of candidate answers for task u. In our implementation, $\mathcal{Y}$ is constructed by combining ground-truth answers with a set of model-generated candidate answers, followed by deduplication. We use $y \in \mathcal{Y}$ to denote a candidate answer.

We compute answer probabilities by conditioning the language model on the task u, the injected evidence, and a fixed reasoning trace c, which serves as a deterministic representation of the agent's current reasoning state. Note that we isolate the effect of evidence by conditioning only on the injected evidence and the reasoning trace, rather than the full state. Concretely, define the injected evidence accumulated up to the end of the l-th search step as $\mathcal{C}_{t_{l+1}-1}$, and compute a length-normalized answer score

$$\log \tilde{\mathbb{P}}(y \mid u, \mathcal{C}_{t_{l+1}-1}, c) = \frac{1}{|y|} \sum_{i=1}^{|y|} \log \mathbb{P}(y_i \mid y_{<i}, u, \mathcal{C}_{t_{l+1}-1}, c), \tag{10}$$

where $|y|$ denotes the length of the answer sequence. We exponentiate and normalize over the candidate set to obtain a normalized answer distribution:

$$\mathbb{P}_l(y) = \frac{\tilde{\mathbb{P}}(y \mid u, \mathcal{C}_{t_{l+1}-1}, c)}{\sum_{y' \in \mathcal{Y}} \tilde{\mathbb{P}}(y' \mid u, \mathcal{C}_{t_{l+1}-1}, c)}, \qquad y \in \mathcal{Y}. \tag{11}$$



(a) Novelty (b) Effectiveness (c) Utility

Figure 6: Per-rollout information novelty, effectiveness and utility across search steps. Each curve corresponds to a rollout, illustrating how novelty, effectiveness and utility evolve with additional evidence.

We generate c using deterministic decoding under each evidence condition, which isolates the effect of newly injected information from stochastic variations in reasoning paths. Effectiveness is defined as the magnitude of change in the model's answer belief induced by the injected evidence contributed by the l-th search step. Specifically, we measure the total variation distance between consecutive answer distributions:

$$\text{Effectiveness}(\Delta \mathcal{C}_l \mid u, s_{t_l}) = \sum_{y \in \mathcal{Y}} \frac{1}{2} \left| \mathbb{P}_l(y) - \mathbb{P}_{l-1}(y) \right|. \tag{12}$$

The effectiveness is bounded in $[0, 1]$. By construction, effectiveness is high when newly injected information substantially alters the model's confidence over candidate answers, and low when additional injected information does not meaningfully affect answer beliefs.

We illustrate novelty, effectiveness, and utility evolve with additional evidence in Figure 6.

Properties. The proposed information utility satisfies the following intuitive properties under our definitions:

1) **Monotonicity with novel and useful evidence.** When newly retrieved evidence provides information that is both novel with respect to the current reasoning state and effective for the task, the information utility increases accordingly. Conversely, evidence that is redundant or task-irrelevant yields little utility gain.
2) **Diminishing returns after task completion.** After sufficient evidence for solving the task has been acquired, additional retrievals tend to contribute increasingly redundant or low-impact information, leading to diminishing marginal utility.

Theoretical analysis under the adopted utility formulation are provided in Section B.

Discussion. We use information utility as an **external control signal**, rather than incorporating it directly into the RL reward. This design choice is motivated by: 1) decoupling agent policy learning from utility estimation makes the framework more general (the utility can be modified without retraining the agent); 2) optimizing the agent policy primarily for process and outcome correctness empirically leads to simpler and more stable RL training.

B THEORETICAL ANALYSIS

B.1 ASSUMPTIONS

The following assumptions are standard in retrieval-augmented reasoning settings and are empirically observed in our experiments.

(A1) **Answer belief convergence.** After sufficient task-relevant evidence has been incorporated into the context, the normalized answer distribution converges, i.e.,

$$\mathbb{P}_l(y) \approx \mathbb{P}_{l-1}(y), \qquad \forall y \in \mathcal{Y}, \tag{13}$$

for sufficiently large l.

(A2) **Evidence redundancy accumulation.** As the number of search steps increases, newly retrieved evidence become increasingly similar, on average, to previously observed evidence, due to the finite amount of task-relevant information.

These assumptions reflect the fact that real-world information spaces contain limited task-relevant content and that repeated retrieval tends to surface increasingly redundant evidence.

B.2 MONOTONICITY WITH NOVEL AND USEFUL EVIDENCE

Lemma 1 (Monotonicity). *If a search step introduces evidence that is both novel and effective, then the information utility is strictly positive.*

Proof. By definition, both $\text{Novelty}(e_l \mid s_{t_l})$ and $\text{Effectiveness}(e_l \mid u, s_{t_l})$ are non-negative and bounded in $[0, 1]$. If

$$\text{Novelty}(e_l \mid s_{t_l}) > 0 \quad \text{and} \quad \text{Effectiveness}(e_l \mid u, s_{t_l}) > 0,$$

and $\rho \in (0, 1)$, then their convex combination satisfies

$$U(e_l \mid u, s_{t_l}) = \rho \cdot \text{Novelty}(e_l \mid s_{t_l}) + (1 - \rho) \cdot \text{Effectiveness}(e_l \mid u, s_{t_l}) > 0. \tag{14}$$

□

Lemma 1 formalizes the intuition that information utility increases only when newly retrieved evidence introduces content that is both non-redundant with respect to the current reasoning state and useful for the downstream task.

B.3 DIMINISHING RETURNS AFTER TASK COMPLETION

Lemma 2 (Diminishing Returns). *Under Assumptions (A1) and (A2), the marginal information utility vanishes asymptotically as the number of search steps increases.*

Proof. We analyze the two components of the utility separately.

Effectiveness. By Assumption (A1), the normalized answer distribution $\mathbb{P}_l(\cdot)$ converges after sufficient task-relevant evidence has been incorporated. Since effectiveness is defined as the total variation distance between $\mathbb{P}_l$ and $\mathbb{P}_{l-1}$, it follows that

$$\lim_{l \to \infty} \text{Effectiveness}(e_l \mid u, s_{t_l}) = 0. \tag{15}$$

Novelty. By Assumption (A2), newly retrieved evidence become increasingly similar to previously observed evidence on average. Since novelty is defined as one minus the average k-nearest-neighbor cosine similarity, this implies

$$\lim_{l \to \infty} \text{Novelty}(e_l \mid s_{t_l}) = 0. \tag{16}$$

Combining the two limits and using the linearity of the utility definition, we obtain

$$\lim_{l \to \infty} U(e_l \mid u, s_{t_l}) = 0. \tag{17}$$

□

Lemma 2 captures diminishing returns in an asymptotic sense: after sufficient task-relevant information has been acquired, additional retrievals tend to introduce increasingly redundant or low-impact evidence, yielding vanishing marginal utility.

C ADAPTIVE INFORMATION CONTROL

C.1 GRANULARITY CONTROL VIA HIERARCHICAL SELECTIVE EXPANSION

In real-world settings, retrieved information can be voluminous and lengthy, making it computationally expensive and often impractical to inject all retrieved content into the agent context. Moreover, fine-grained details are not uniformly useful across reasoning stages. We therefore introduce *granularity control*, which presents retrieval results at a coarse level first and allows the agent to selectively expand higher-granularity information only when needed.

Under granularity control, **retrieval and information refinement are decoupled**: the agent first retrieves coarse-grained information via `retrieve`, and then selectively refines it through explicit `expand` actions. Formally, we model external information as a hierarchical structure (Figure 3). At search step l, the search engine returns a set of k sources $e_l = \{\mathcal{G}_l^{(i)}\}_{i=1}^{k}$, where each source is represented as a rooted tree $\mathcal{G}_l^{(i)} = (\mathcal{V}_l^{(i)}, \mathcal{E}_l^{(i)})$. Each node $v \in \mathcal{V}_l^{(i)}$ corresponds to an evidence unit at a particular resolution, and each directed edge $(v, v') \in \mathcal{E}_l^{(i)}$ indicates that v' is a refinement of v.

After retrieval, instead of injecting all leaf-level content, we initialize by appending only the retrieved root nodes to the current context; the resulting injected set is denoted by $\mathcal{C}_{t_l}$. The agent may then perform a variable number of `expand` actions to incrementally grow the observed set until the next retrieval (or termination). Let t_{l+1} denote the primitive step of the next retrieval (or termination boundary), so that all expansions triggered by the l-th retrieval are completed by step $t_{l+1} - 1$. The injected nodes satisfy $\mathcal{C}_{t_l} \subseteq \mathcal{C}_{t_l+1} \subseteq \cdots \subseteq \mathcal{C}_{t_{l+1}-1} \subseteq \bigcup_{i=1}^{k} \mathcal{V}_l^{(i)}$, and are expanded adaptively as needed. The net injected nodes contributed by search step l are $\Delta\mathcal{C}_l = \mathcal{C}_{t_{l+1}-1} \setminus \mathcal{C}_{t_l-1}$, where $\mathcal{C}_{t_l-1}$ is the injected set right before the l-th retrieval.

For $t' \in \{t_l + 1, \ldots, t_{l+1} - 1\}$, an expansion action at primitive step t' is defined as $a_{t'} = (h_{t'}, \alpha_{t'}, \xi_{t'})$, where $h_{t'}$ denotes the agent's thought, $\alpha_{t'} = \texttt{expand}$, and the action parameters $\xi_{t'} \subseteq \bigcup_{i=1}^{k} \mathcal{E}_l^{(i)}$ specify a set of hierarchy edges (v, v') such that $v \in \mathcal{C}_{t'-1}$ and v' is a child of v in the corresponding tree. Executing $a_{t'}$ updates $\mathcal{C}_{t'} = \mathcal{C}_{t'-1} \cup \{ v' \mid v \in \mathcal{C}_{t'-1}, (v, v') \in \xi_{t'} \}$, i.e., newly expanded nodes are added to the observed evidence set.

During training, given the retrieved hierarchies e_l, we derive the expansion targets $\{\mathcal{C}^*_{t_l+1}, \ldots, \mathcal{C}^*_{t_{l+1}-1}\}$ using the information utility signal $U(\cdot)$, and use them to guide the agent's expansion decisions. Concretely, we score all leaf nodes in the retrieved trees and select the top-k_{expand} leaves. We then trace these leaves upward, collecting their ancestors layer by layer until reaching the root, which yields the target observed evidence sets $\{\mathcal{C}^*_{t_l+1}, \ldots, \mathcal{C}^*_{t_{l+1}-1}\}$. Given this target, the controller provides explicit guidance in the form of desired expansion edges $\xi^*_{t'}$ for $t' \in \{t_l + 1, \ldots, t_{l+1} - 1\}$, so that the induced updates follow

$$\mathcal{C}^*_{t'} = \mathcal{C}^*_{t'-1} \cup \{ v' \mid v \in \mathcal{C}^*_{t'-1}, (v, v') \in \xi^*_{t'} \}. \tag{18}$$

The model is trained to select expansion actions aligned with $\xi^*_{t'}$, thereby learning a granularity-control policy that prioritizes high-utility information while minimizing context growth.

C.2 SEARCH CONTINUATION CONTROL

By default, the agent autonomously decides whether to search based on its internal reasoning state. However, this decision is often suboptimal: the agent may terminate search prematurely by underestimating the value of additional information, or overcommit to continued search when no further useful evidence is available. We therefore model *search continuation* as an explicit control decision, where external intervention is applied *only* when utility signals indicate systematic misjudgment (Figure 2).

Termination. If the information utility remains below a threshold δ for m_{stop} consecutive search steps, we define the stopping index

$$l^\star = \min_{l \in [m_{\text{stop}}-1, L-1]} \max_{j \in [l-m_{\text{stop}}+1, l]} U(e_j) < \delta. \tag{19}$$

Upon reaching $l^\star$, a control signal $\kappa = \texttt{Stop searching}$ is injected, explicitly terminating further search steps.

Continuation. Conversely, the agent may attempt to terminate search and proceed to answer generation even when additional evidence is still beneficial. Let $\mathbb{P}_l(y)$ denote a model-based probability for the candidate answer y (defined in Equation (11)). If the agent attempts to terminate search at index l but the draft answer probability is low, while the utility of the most recent m_{cont} searches remains consistently high, we trigger an intervention to force one additional search step:

$$\mathbb{P}_l(y_{\text{gold}}) < \eta \cdot \max_{y \in \mathcal{Y}} \mathbb{P}_l(y) \;\wedge\; \min_{j \in [l-m_{\text{cont}}+1, l]} U(e_j) \geq \delta, \tag{20}$$

where η is the probability threshold. Upon satisfying Equation (20), a one-shot control signal $\kappa = \texttt{Continue the search for one additional step}$ is injected. Note that Equation (20) is used only during *training* when y_{gold} is available.

Discussion. Under this setting, search continuation is primarily governed by the agent's learned policy, while information utility serves as a monitoring signal that triggers corrective control when necessary. Detailed hyperparameter settings and ablations are provided in Section D

C.3 REINFORCEMENT LEARNING WITH INFORMATION CONTROL

Agents can use *external control signals* to improve exploration and stabilize early-stage learning (Figure 4), but the acquired strategies need be internalized into model parameters to enhance intrinsic capabilities at test time. To this end, we propose two rollout modes under an annealed *control-forcing* RL scheme, and introduce a composite reward that combines answer correctness, tool-usage regularization, and retrieval effectiveness.

Rollout Modes. During rollouts, the agent samples between two modes, selecting mode (1) with probability p and mode (2) with $1 - p$.

(1) With Information Control. For each task u, a controller monitors the utility of retrieved information throughout the rollout. Upon detecting an abnormal retrieval pattern, the controller triggers a control signal κ at time $t^\star$. Conditioned on the current reasoning state $s_{t^\star}$ and the triggered control signal κ, the policy generates the next action as $a_{t^\star} \sim \pi_\theta(\cdot \mid u, s_{t^\star}, \kappa)$.

(2) Without Information Control. For each task u, at each step t, the policy π_θ generates thoughts and actions conditioned only on the current state s_t and task: $a_t \sim \pi_\theta(\cdot \mid u, s_t)$.

The prompts corresponding to the two rollout modes are provided in Section D.

Update Modes. We adopt an annealed *control-forcing curriculum* that gradually removes control signals so that the final policy performs reliably without external intervention. Concretely, we schedule p across epochs and optimize under a progressively shifting mixture of the two rollout modes: early training uses frequent control, mid training reduces control, and the final stage removes control entirely. Within each stage, rollouts are generated by the current policy under the corresponding observation regime (i.e., the control signal, when present, is included in the context), and we perform *on-policy* updates with respect to that regime. Compared with vanilla RL, this curriculum improves stability in early training when the agent is not yet able to produce effective rollouts without guidance, while ensuring that the learned behavior transfers to the no-control setting at convergence.

Reward Design. For online RL, reward design is critical, as the learning process is directly driven by reward signals. Motivated by this property, we design a composite reward that integrates answer correctness, tool-usage regularization, and retrieval effectiveness, providing informative learning signals for search behavior while preserving an outcome-driven reinforcement learning objective. Building upon outcome rewards based on F1 score, we incorporate explicit penalties for improper tool usage and a bonus for effective retrieval. The final reward for a reasoning trajectory τ is defined as

$$\hat{r}_\phi(\tau, y_{\text{gold}}) \;=\; r_{\text{correct}}(\tau, y_{\text{gold}}) + r_{\text{penalty}}(\tau) + r_{\text{ret}}(\tau, y_{\text{gold}}). \tag{21}$$

where y_{gold} is the gold answer, and ϕ denotes reward hyperparameters.

The base reward of correctness is

$$r_{\text{correct}}(\tau, y_{\text{gold}}) \;=\; \begin{cases} \max\big(\mathrm{F1}(y_{\text{pred}}, y_{\text{gold}}),\ \lambda_{\text{format}}\big), & y_{\text{pred}} \neq \varnothing, \\ 0, & \text{otherwise}, \end{cases} \tag{22}$$

where λ_{format} is a format floor ensuring that valid outputs receive a non-zero reward.

To discourage improper tool interactions, we introduce a tool-usage penalty

$$r_{\text{penalty}}(\tau) \;=\; -\min\big(\lambda_{\text{penalty}} \cdot N_{\text{penalty}}(\tau),\ \lambda_{\text{penalty}}^{\max}\big), \tag{23}$$

where $N_{\text{penalty}}(\tau)$ counts the number of tool-usage violations in the trajectory. We consider two types of violations: (i) incorrect tool usage, such as issuing malformed inputs; and (ii) control non-compliance, where the agent fails to follow explicit control messages. Each violation incurs a penalty scaled by λ_{penalty}, with the total penalty capped at $\lambda_{\text{penalty}}^{\max}$ to avoid over-penalization.

Finally, we include a retrieval bonus

$$r_{\text{ret}}(\tau, y_{\text{gold}}) \;=\; \begin{cases} \lambda_{\text{ret}}, & \mathbb{I}_{\text{ret}}(\tau, y_{\text{gold}}) = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

where $\mathbb{I}_{\text{ret}}(\tau, y_{\text{gold}})$ indicates whether the retrieved documents contain the ground-truth answer, instantiated in our experiments via substring exact match between retrieved passages and the target answer.

The total reward is further capped by an imperfect ceiling λ_{ceil}, preventing trajectories with incorrect final answers from receiving maximal reward.

$$r_\phi(\tau, y_{\text{gold}}) \; = \; \begin{cases} \min(\hat{r}_\phi(\tau, y_{\text{gold}}), 1), & \text{F1}(y_{\text{pred}}, y_{\text{gold}}) = 1, \\ \min(\hat{r}_\phi(\tau, y_{\text{gold}}), \lambda_{\text{ceil}}), & \text{otherwise}, \end{cases} \tag{25}$$

Detailed hyperparameter settings and ablations are provided in Section D.

D IMPLEMENTATION DETAILS

Prompts. Below, we present all prompts used in our framework, including the search-augmented reasoning prompt, the control message, and the candidate answer generation prompt.

Search-Augmented Reasoning Prompt

```
[SYSTEM PROMPT]
You are a helpful assistant.

[USER PROMPT]
Answer the given question.

You MUST follow the protocol below.

CONTROL
- A control message may appear anywhere in the conversation in the form:
  <control>...</control>
- You MUST follow the <control> message that appears in the context.


General rules
- Whenever you receive NEW information (from <search_results>, <information>), you MUST
  first reason inside <think>...</think>.
- You can call a search engine using: <search>query</search>.
  The environment will return snippets inside: <search_results>...</search_results>.
- If you want full text, you MUST decide inside <think>...</think>, then request expansion
  using: <expand>{"doc_ids": [id1, id2, ...]}</expand>
  The environment will return the expanded full text inside: <information>...</info-
  rmation>. You can expand multiple documents in one call by listing multiple doc_ids.
- If no further external knowledge is needed, output the final answer inside <answer>...
  </answer>.


Answer normalization rules (VERY IMPORTANT)
- The final answer MUST EXACTLY match the canonical short answer.
- Output the SHORTEST possible answer span.
- Do NOT add explanations, appositives, or parentheses.
- Do NOT add extra words, punctuation, or formatting.
- Use the most common name form that appears as a standalone answer.
- If multiple aliases exist, choose the most standard short form.
- Case-sensitive matching is required.

Examples:
Q: how many episodes are in series 7 game of thrones?
Correct: <answer>seven</answer>

Q: when does season 5 of bates motel come out?
Correct: <answer>February 20, 2017</answer>


Round definition
A round MUST be one of the following two sequences:

1) Answering round:
   <think>...</think>
   <search>...</search>
   <search_results>...</search_results>
   <think>...</think>
   <expand>...</expand>
   <information>...</information>
   <think>...</think>
```

```
    <answer>...</answer>

2) Continuing round:
    <think>...</think>
    <search>...</search>
    <search_results>...</search_results>
    <think>...</think>
    <expand>...</expand>
    <information>...</information>
    <think>...</think>

You may perform as many rounds as needed.

Question: {question}
```

Control Message

```
<control>Expand the retrieved documents: [...]</control>

<control>Stop searching</control>

<control>Continue the search for one additional step</control>
```

Candidate Answer Generation Prompt

```
**Task:**

You will be given a **question**. Write some plausible candidate answers to the question.

Example 1:
**Input:**
Q: The Oberoi family is part of a hotel company that has a head office in what city?

**Output:**
["Delhi", "Shanghai", "New York", "Madrid", "Bangkok"]


Example 2:
**Input:**
Q: Which magazine was started first Arthur's Magazine or First for Women?

**Output:**
["Arthur's Magazine", "First for Women", "Both magazines were started in the same year."]


Example 3:
**Input:**
Q: What nationality was James Henry Miller's wife?

**Output:**
["Scottish", "English", "Irish", "Welsh", "American", "Japanese", "Spanish", "Korean",
"Canadian"]


Test:
**Input:**
Q: {question}

**Output:**
```

Control Hyperparameters. Unless otherwise specified, we use a unified set of control hyperparameters across all tasks. For novelty estimation, we set the number of nearest neighbors to $k_{\text{nn}} = 5$. The utility weight $\rho = 0.5$ balances the contributions of novelty and effectiveness. The utility threshold is set to $\delta = 0.2$, which corresponds to a conservative cutoff below which newly retrieved information exhibits limited marginal value. For termination control, we use a short window $m_{\text{stop}} = 2$, which we find sufficient to robustly detect sustained low-utility search steps due to the stability of search-level utility signals. For continuation control, we adopt a one-step window $m_{\text{cont}} = 1$ to promptly correct premature stopping based on recent high-utility evidence. The answer likelihood threshold is set to $\eta = 0.7$, applied to length-normalized and normalized answer probabilities. We observe that the overall control behavior is insensitive to small variations around these values.

Reward Hyperparameters. Unless otherwise specified, we use a fixed set of reward hyperparameters across all tasks. The format floor is set to $\lambda_{\text{format}} = 0.1$, ensuring that trajectories producing validly formatted outputs receive a minimal positive signal, which stabilizes early-stage training without overshadowing answer correctness. The per-violation tool-usage penalty is set to $\lambda_{\text{penalty}} = 0.2$, with the maximum penalty capped at $\lambda_{\text{penalty}}^{\max} = 0.4$, preventing excessive penalization from dominating the reward signal in trajectories with multiple violations. The retrieval bonus is set to $\lambda_{\text{ret}} = 0.1$, providing a mild incentive for retrieving documents that contain the ground-truth answer, while avoiding over-reliance on retrieval signals. Finally, the imperfect reward ceiling is set to $\lambda_{\text{ceil}} = 0.9$, ensuring that trajectories with incorrect final answers cannot achieve maximal reward, even when other auxiliary signals are favorable. We find training to be robust to moderate variations of these values.

Training Setup. We conduct experiments with Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct (Yang et al., 2024). For retrieval, we use the 2018 Wikipedia dump (Karpukhin et al., 2020) as the knowledge source and E5 (Wang et al., 2022) as the retriever. Unlike prior methods that append raw retrieved passages to the context, our approach uses hierarchical selective expansion. For fair comparison, following (Lin et al., 2023), we set the number of retrieved passages to 3 for all existing retrieval-based baselines. For our method, we retrieve 5 candidate summaries but cap evidence usage by limiting the agent to at most 3 expansion nodes, matching the effective evidence budget.

For training, following (Jin et al., 2025), we merge the training sets of NQ and HotpotQA to form a unified dataset for DEEPCONTROL. We adopt PPO as the RL algorithm, as we observed that GRPO leads to training collapse after a few dozen of optimization steps. We train for 5 epochs in total and anneal the control probability p in stages, using $p = 0.9$, 0.5, 0.2, and 0 for 2, 1, 1, and 1 epochs, respectively. Evaluation is conducted on the test or validation sets of seven datasets to assess both in-domain and out-of-domain performance. Exact Match (EM) is used as the evaluation metric, following Yu et al. (2024). For inference-style baselines, we use instruct models, as base models fail to follow instructions. For RL tuning methods, experiments are conducted on both base and instruct models.

For the PPO variant of DEEPCONTROL, we follow the implementation provided in Verl (Sheng et al., 2024) and set the learning rate of the policy model to 1×10^{-6} and that of the value model to 1×10^{-5}. Training is performed with warm-up ratios of 0.1 and 0.015 for the policy and value models, respectively. We employ Proximal Policy Optimization with Generalized Advantage Estimation (GAE), using $\lambda_{\text{GAE}} = 1$ and $\gamma_{\text{GAE}} = 1$.

All PPO experiments are conducted on a single node equipped with eight A100 GPUs. We use a training batch size of 64 per update, with a PPO mini-batch size of 64 and a micro-batch size of 4 for both the policy and value networks. The maximum prompt length is set to 5,120 tokens, with a maximum response length of 512 tokens. To reduce GPU memory consumption, we enable gradient checkpointing and employ Fully Sharded Data Parallel (FSDP) training with CPU parameter offloading.

For efficient rollout generation, we adopt vLLM (Kwon et al., 2023) with a tensor parallel size of 1 and a GPU memory utilization ratio of 0.4. Rollout sampling uses a temperature of 1.0. We use an adaptive KL controller with an initial coefficient of $\beta = 0.001$, together with standard PPO clipping.

For GRPO training, we set the policy learning rate to 1×10^{-6}. We sample six responses per prompt and train the model with a warm-up ratio of 0.1. GRPO experiments are conducted using the same hardware setup, a training batch size of 32, sequence length limits, and rollout configurations as in PPO. We use a larger explicit KL penalty ($\beta = 0.01$) for improved training stability. Unless otherwise specified, gradient checkpointing, FSDP offloading, and vLLM-based rollouts share identical hyperparameters across methods.

Model checkpoints are saved every 100 training steps. If training becomes unstable, we select the most recent stable checkpoint based on the reward curve; otherwise, the final checkpoint is used for evaluation. Unless stated otherwise, we set the maximum action budget to 8. PPO is used as the default RL algorithm, with a detailed comparison between PPO and GRPO provided in Section E. All experiments are conducted with a fixed random seed.

E ADDITIONAL RESULTS

DEEPCONTROL vs. PPO. We compare DEEPCONTROL against vanilla PPO without control signals. Both methods are trained using the same data, reward design, and hyperparameter configuration. The training dynamics are shown in Figure 5(a), and the evaluation results are reported in Table 3. DEEPCONTROL consistently achieves higher performance than vanilla PPO. The control signals provide corrective guidance during early training, helping the agent avoid suboptimal retrieval behaviors when the policy is still immature. As training progresses, these behaviors are gradually internalized by the policy, allowing the agent to perform effectively even after control signals are removed. On average, DEEPCONTROL improves performance by 8.3% over vanilla PPO, demonstrating that information control substantially enhances learning efficiency in online RL.

Table 3: Comparison of our method with vanilla PPO.

Method	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	Avg.
Qwen2.5-3b-Instruct								
DEEPCONTROL	**0.533**	**0.645**	**0.512**	**0.402**	**0.371**	**0.118**	**0.298**	**0.411**
PPO	0.432	0.518	0.413	0.307	0.293	0.094	0.237	0.328

PPO vs. GRPO. We evaluate DEEPCONTROL using PPO and GRPO as the underlying RL algorithm. The training dynamics are shown in Figure 5(b), and the final results are summarized in Table 4. We observed that (1) GRPO converges faster than PPO in early training. This behavior is expected, as PPO relies on a learned critic, which typically requires a warm-up period before providing reliable value estimates. (2) PPO exhibits greater stability under control annealing. As shown in Figure 5(b), GRPO suffers from reward collapse after extended training, whereas PPO maintains stable optimization throughout the annealing process. (3) PPO achieves higher final performance than GRPO. Due to reward collapse under annealed control, policies trained with GRPO perform worse than PPO when evaluated without control signals, highlighting PPO's robustness in this setting.

Table 4: Comparison of our method implemented with PPO and GRPO.

Method	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	Avg.
Qwen2.5-3b-Instruct								
DEEPCONTROL (PPO)	**0.533**	**0.645**	**0.512**	**0.402**	**0.371**	**0.118**	**0.298**	**0.411**
DEEPCONTROL (GRPO)	0.362	0.438	0.348	0.271	0.254	0.081	0.202	0.279

Example outputs. Below, we present representative examples of DEEPCONTROL under settings with and without the control signal, evaluated on both single-hop and multi-hop questions.

Search-Augmented Reasoning – NQ (no control)

```
[SYSTEM PROMPT]
You are a helpful assistant.

[USER PROMPT]
Answer the given question.

You MUST follow the protocol below.

CONTROL
- A control message may appear anywhere in the conversation in the form:
  <control>...</control>
- You MUST follow the <control> message that appears in the context.


General rules
- Whenever you receive NEW information (from <search_results>, <information>), you MUST
  first reason inside <think>...</think>.
- You can call a search engine using: <search>query</search>.
  The environment will return snippets inside: <search_results>...</search_results>.
- If you want full text, you MUST decide inside <think>...</think>, then request expansion
```

```
  using: <expand>{"doc_ids": [id1, id2, ...]}</expand>
  The environment will return the expanded full text inside: <information>...</info-
  rmation>. You can expand multiple documents in one call by listing multiple doc_ids.
- If no further external knowledge is needed, output the final answer inside <answer>...
  </answer>.


Answer normalization rules (VERY IMPORTANT)
- The final answer MUST EXACTLY match the canonical short answer.
- Output the SHORTEST possible answer span.
- Do NOT add explanations, appositives, or parentheses.
- Do NOT add extra words, punctuation, or formatting.
- Use the most common name form that appears as a standalone answer.
- If multiple aliases exist, choose the most standard short form.
- Case-sensitive matching is required.

Examples:
Q: how many episodes are in series 7 game of thrones?
Correct: <answer>seven</answer>

Q: when does season 5 of bates motel come out?
Correct: <answer>February 20, 2017</answer>


Round definition
A round MUST be one of the following two sequences:

1) Answering round:
   <think>...</think>
   <search>...</search>
   <search_results>...</search_results>
   <think>...</think>
   <expand>...</expand>
   <information>...</information>
   <think>...</think>
   <answer>...</answer>

2) Continuing round:
   <think>...</think>
   <search>...</search>
   <search_results>...</search_results>
   <think>...</think>
   <expand>...</expand>
   <information>...</information>
   <think>...</think>

You may perform as many rounds as needed.

Question: when was the first election held in india?


[ASSITANT RESPONSE]
<think>To answer the question, I will first perform a search about The First General
Election of India.</think> <search>The First General Election of India</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=21045267 score=0.8642
Title: First General Election of the Republic
Snippet: The first general election to the Lok Sabha, conducted after India adopted a
federal structure and a constitutional framework that enshrined universal adult suffrage,
stands as a defining moment in modern Indian democrac...

Rank=2 doc_id=21074657 score=0.8551
Title: India's First General Election
Snippet: The first general elections held after independence represented a milestone in
the consolidation of democratic practice in the subcontinent, adapting the newly framing
constitutional framework to a nationwide electoral s...

Rank=3 doc_id=21097275 score=0.8536
Title: First General Elections in India
Snippet: In the wake of a hard-won independence and the adoption of a democratic
constitution, the Indian republic undertook a comprehensive electoral process to translate
sovereign ideals into executive reality. ...

Rank=4 doc_id=21025381 score=0.8509
Title: First General Election in India
Snippet: The transfer from colony to republic culminated in a nationwide electoral
```

```
exercise that tested the logistical and democratic capacities of a vast, diverse
electorate, spanning hundreds of languages, cultures, and rural-u...

Rank=5 doc_id=17180261 score=0.8472
Title: 1951{52 elections in India
Snippet: 1951{52 elections in India The Election Commission of India held the first
presidential elections of India on May 2, 1952. ...
</search_results>


[ASSITANT RESPONSE]
<think>The search results show documents about The First General Election of India, but
the snippets don't give the specific answer. I need to expand the most relevant documents
to find when was the first election held in india. Document 21074657 seems most promising.
</think>
<expand>{"doc_ids": [21074657]}</expand>


[USER PROMPT]
<information>
[DOC 21074657]
"India's First General Election"

The first general elections held after independence represented a milestone in the
consolidation of democratic practice in the subcontinent, adapting the newly framing
constitutional framework to a nationwide electoral system. Under the Representation of the
People Act, enacted in 1950, and the provisions of the newly adopted Constitution,
universal adult suffrage was implemented for all citizens, irrespective of gender, caste,
or religion, and a complex logistical undertaking was mounted to administer campaigns,
voter registration, and polling across thousands of constituencies. The process unfolded
over a extended campaign year that spanned two calendar years as electoral rolls were
updated, constituencies delineated, and the machinery of government prepared for peaceful
transfer of power through regular competitive elections. The culmination of these efforts
is recorded in historical summaries as taking place during 1951{52, a period during which
voting proceeded in multiple phases across states and union territories, with results
announced progressively and inaugurations of elected representatives following thereafter.
This electoral event established the template for subsequent national elections, embedding
the practice of periodic, inclusive suffrage within the constitutional life of the
republic and shaping path dependencies for party organization, coalition-building, and
regional representation in the decades that followed.
</information>


[ASSITANT RESPONSE]
<think>The paragraph states the electoral event took place during 1951{52.</think>
<answer>1951{52</answer>
```

Search-Augmented Reasoning – HotpotQA (no control)

```
[SYSTEM PROMPT]
You are a helpful assistant.

[USER PROMPT]
Answer the given question.

You MUST follow the protocol below.

CONTROL
- A control message may appear anywhere in the conversation in the form:
  <control>...</control>
- You MUST follow the <control> message that appears in the context.


General rules
- Whenever you receive NEW information (from <search_results>, <information>), you MUST
  first reason inside <think>...</think>.
- You can call a search engine using: <search>query</search>.
  The environment will return snippets inside: <search_results>...</search_results>.
- If you want full text, you MUST decide inside <think>...</think>, then request expansion
  using: <expand>{"doc_ids": [id1, id2, ...]}</expand>
  The environment will return the expanded full text inside: <information>...</info-
  rmation>. You can expand multiple documents in one call by listing multiple doc_ids.
- If no further external knowledge is needed, output the final answer inside <answer>...
  </answer>.


Answer normalization rules (VERY IMPORTANT)
```

```
- The final answer MUST EXACTLY match the canonical short answer.
- Output the SHORTEST possible answer span.
- Do NOT add explanations, appositives, or parentheses.
- Do NOT add extra words, punctuation, or formatting.
- Use the most common name form that appears as a standalone answer.
- If multiple aliases exist, choose the most standard short form.
- Case-sensitive matching is required.

Examples:
Q: how many episodes are in series 7 game of thrones?
Correct: <answer>seven</answer>

Q: when does season 5 of bates motel come out?
Correct: <answer>February 20, 2017</answer>


Round definition
A round MUST be one of the following two sequences:

1) Answering round:
   <think>...</think>
   <search>...</search>
   <search_results>...</search_results>
   <think>...</think>
   <expand>...</expand>
   <information>...</information>
   <think>...</think>
   <answer>...</answer>

2) Continuing round:
   <think>...</think>
   <search>...</search>
   <search_results>...</search_results>
   <think>...</think>
   <expand>...</expand>
   <information>...</information>
   <think>...</think>

You may perform as many rounds as needed.

Question: What UNESCO World Heritage Site is located in Tamar Ziskind’s hometown?


[ASSITANT RESPONSE]
<think>To answer the question, I will first perform a search about Heritage Sites and the
Mediterranean Coast.</think>
<search>Heritage Sites and the Mediterranean Coast</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=20132687 score=0.8224
Title: Heritage coast
Snippet: Heritage coast A heritage coast is a strip of coastline in England and Wales, the
extent of which is defined by agreement between the relevant statutory national agency and
the relevant local authority. ...

Rank=2 doc_id=21505740 score=0.8176
Title: National heritage site
Snippet: A national heritage site is a heritage site having a value that has been
registered by a governmental agency as being of national importance to the cultural
heritage or history of that country. ...

Rank=3 doc_id=4178593 score=0.8125
Title: World Heritage Site
Snippet: World Heritage Site A World Heritage Site is a landmark or area which is selected
by the United Nations Educational, Scientific and Cultural Organization (UNESCO) as having
cultural, historical, scientific or other form ...

Rank=4 doc_id=18702191 score=0.8108
Title: Heritage coast
Snippet: has only very sporadic stretches. The first heritage coast was Beachy Head with
its famous white cliffs. Heritage coasts listed clockwise around the English coast from
Northumberland: Heritage coasts listed clockwise aro...

Rank=5 doc_id=10719879 score=0.8094
Title: Causses and Cévennes
Snippet: the impact thereof on national history. In 2009, ICOMOS also requested that
France justify the boundary of the property and provide a basis for managing and
```

```
sustaining the site’s agro-pastoral traditions. ...
</search_results>


[ASSITANT RESPONSE]
<think>The first result is about Heritage coast but the snippet doesn’t mention Tamar
Ziskind or Haifa. I need to expand it to see the full content and find the relevant
information.</think>
<expand>{"doc_ids": [20132687]}</expand>


[USER PROMPT]
<information>
[DOC 20132687]
"Heritage coast"
Heritage coast A heritage coast is a strip of coastline in England and Wales, the extent
of which is defined by agreement between the relevant statutory national agency and the
relevant local authority. Such areas are recognised for their natural beauty, wildlife and
heritage and amongst the purposes of definition is support for these qualities and
enabling enjoyment of them by the public. For England this national agency is Natural
England (having formerly been the Countryside Agency) and for Wales it is Natural
Resources Wales (which took over the role from its predecessor body, Countryside Council
for Wales). 1,057 km
</information>


[ASSITANT RESPONSE]
<think>The document is about Heritage coast but still doesn’t mention Tamar Ziskind or
Haifa. I need to search more specifically.</think>
<search>Tamar Ziskind Haifa UNESCO World Heritage Site</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=21451362 score=0.8549
Title: Tamar Ziskind
Snippet: Tamar Ziskind (Hebrew: ;born February 23, 1985 in Haifa, Israel) is a beauty
pageant contestant who represented Israel in Miss World 2008 in South Africa. ...

Rank=2 doc_id=13745445 score=0.8266
Title: Haifa
Snippet: Haifa Haifa ( ’ ; ’) is the third-largest city in Israel { after Jerusalem and
Tel Aviv{ with a population of in . ...

Rank=3 doc_id=9256800 score=0.8197
Title: Tourism in Israel
Snippet: BC. Anati instead places the Exodus, based on other archaeological evidence at
around 2300 BC Beit Guvrin-Maresha National Park is a national park in central Israel, 13
kilometers from Kiryat Gat, encompassing the ruins ...

Rank=4 doc_id=20479906 score=0.8188
Title: Ein Tamar
Snippet: Ein Tamar earn their livelihood from agriculture. Peppers and melons the most
popular crops. Ein Tamar Ein Tamar (, ""lit. ...

Rank=5 doc_id=20479905 score=0.8174
Title: Ein Tamar
Snippet: Ein Tamar Ein Tamar (, ""lit."" Date Palm Spring) is a moshav in the northern
Arava valley in Israel. Located south of the Dead Sea, it falls under the jurisdiction of
Tamar Regional Council. ...
</search_results>


[ASSITANT RESPONSE]
<think>The first result is about Tamar Ziskind. I need to expand it to see the full text
and find the relevant information.</think>
<expand>{"doc_ids": [21451362]}</expand>


[USER PROMPT]
<information>
[DOC 21451362]
"Tamar Ziskind"
Tamar Ziskind (Hebrew: ;born February 23, 1985 in Haifa, Israel) is a beauty pageant
contestant who represented Israel in Miss World 2008 in South Africa. She has worked as a
psychometrics instructor and has studied dentistry.
</information>
```

```
[ASSITANT RESPONSE]
<think>The expanded document still doesn’t mention UNESCO World Heritage Site. I need to
search more broadly for information about UNESCO World Heritage Site in Haifa.</think>
<search>UNESCO World Heritage Site in Haifa beach hill country Israel</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=13745445 score=0.8647
Title: Haifa
Snippet: Haifa Haifa ( ’ ; ’) is the third-largest city in Israel { after Jerusalem and
Tel Aviv{ with a population of in . ...

Rank=2 doc_id=13745498 score=0.8460
Title: Haifa
Snippet: five buildings with 85,000 square meters of space on a 31-duman (7.75 acre) site
is being built adjacent to the Matam industrial park. ...

Rank=3 doc_id=14707018 score=0.8452
Title: Haifa Bay
Snippet: Haifa Bay The Bay of Haifa or Haifa Bay (, ""Mifratz Haifa""), formerly Gulf of
Acre or Bay of Acre, is a bay along the Mediterranean coast of Northern Israel. ...

Rank=4 doc_id=5070278 score=0.8446
Title: Haifa
Snippet: including the Port of Haifa. The middle level is on the slopes of Mount Carmel
and consists of older residential neighborhoods, while the upper level consists of modern
neighborhoods looking over the lower tiers. ...

Rank=5 doc_id=5070277 score=0.8412
Title: Haifa
Snippet: of liberal Arabic
speaking culture, as it was under British colonial rule. The Arabic speaking
neighborhoods, which are mixed Muslim and Christian, are in the lowlands near the sea
while Jewish neighborhoods are at highe...
</search_results>


[ASSITANT RESPONSE]
<think>The search results are not giving me the specific information. The fifth rank is
about Haifa education districts, which might not be relevant. I need to think more about
geography: Haifa is in Israel and has a site. The famous ruins at Tel Aviv are UNESCO
sites. Let me search for known UNESCO World Heritage sites in Haifa.</think>
<search>Tel Aviv UNESCO World Heritage Site Haifa</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=13745445 score=0.8806
Title: Haifa
Snippet: Haifa Haifa ( ’ ; ’) is the third-largest city in Israel { after Jerusalem and
Tel Aviv{ with a population of in . ...

Rank=2 doc_id=21451363 score=0.8532
Title: Haifa
Snippet: Haifa (Hebrew: "Hefa " ] ; Arabic: "Hayfa ") is the third-largest city in the
State of Israel after Jerusalem and Tel Aviv, with a population of 279,591 in 2016 . ...

Rank=3 doc_id=3555477 score=0.8433
Title: Tel Aviv
Snippet: outlook of the Bauhaus and similar schools to the local conditions there,
creating what is recognized as the largest concentration of buildings in the International
Style in the world. ...

Rank=4 doc_id=13745498 score=0.8428
Title: Haifa
Snippet: five buildings with 85,000 square meters of space on a 31-duman (7.75 acre) site
is being built adjacent to the Matam industrial park. ...

Rank=5 doc_id=1346942 score=0.8374
Title: Haifa
Snippet: Haifa is considered a relative haven for coexistence between Jews and Arabs. Two
respected academic institutions, the University of Haifa and the Technion, are located in
Haifa, in addition to the largest k-12 school in ...
</search_results>


[ASSITANT RESPONSE]
<think>The snippet doesn’t show UNESCO information. I need to expand the top document to
```

```
find if it mentions UNESCO sites.</think>
<expand>{"doc_ids": [13745445]}</expand>


[USER PROMPT]
<information>
[DOC 13745445]
Haifa
Haifa Haifa ( ’ ; ’) is the third-largest city in Israel { after Jerusalem and Tel Aviv{
with a population of in . The city of Haifa forms part of the Haifa metropolitan area, the
second- or third-most populous metropolitan area in Israel. It is home to the Bahá’í World
Centre, a UNESCO World Heritage Site and a destination for Bahá’í pilgrims. Built on the
slopes of Mount Carmel, the settlement has a history spanning more than 3,000 years. The
earliest known settlement in the vicinity was Tell Abu Hawam, a small port city
established in the Late Bronze
</information>


[ASSITANT RESPONSE]
<think>The document mentions the Bahá’í World Centre as a UNESCO World Heritage Site. The
question asks: "What UNESCO World Heritage Site is located in Tamar Ziskind’s hometown?"
The hometown is Haifa, Israel. The answer should be the name of the site. The canonical
short answer is "Bahá’í World Centre".</think>
<answer>Bahá’í World Centre</answer>
```

Search-Augmented Reasoning – NQ (with control)

```
[SYSTEM PROMPT]
You are a helpful assistant.

[USER PROMPT]
Answer the given question.

You MUST follow the protocol below.

CONTROL
- A control message may appear anywhere in the conversation in the form:
  <control>...</control>
- You MUST follow the <control> message that appears in the context.


General rules
- Whenever you receive NEW information (from <search_results>, <information>), you MUST
  first reason inside <think>...</think>.
- You can call a search engine using: <search>query</search>.
  The environment will return snippets inside: <search_results>...</search_results>.
- If you want full text, you MUST decide inside <think>...</think>, then request expansion
  using: <expand>{"doc_ids": [id1, id2, ...]}</expand>
  The environment will return the expanded full text inside: <information>...</info-
  rmation>. You can expand multiple documents in one call by listing multiple doc_ids.
- If no further external knowledge is needed, output the final answer inside <answer>...
  </answer>.


Answer normalization rules (VERY IMPORTANT)
- The final answer MUST EXACTLY match the canonical short answer.
- Output the SHORTEST possible answer span.
- Do NOT add explanations, appositives, or parentheses.
- Do NOT add extra words, punctuation, or formatting.
- Use the most common name form that appears as a standalone answer.
- If multiple aliases exist, choose the most standard short form.
- Case-sensitive matching is required.

Examples:
Q: how many episodes are in series 7 game of thrones?
Correct: <answer>seven</answer>

Q: when does season 5 of bates motel come out?
Correct: <answer>February 20, 2017</answer>


Round definition
A round MUST be one of the following two sequences:

1) Answering round:
   <think>...</think>
   <search>...</search>
```

```
    <search_results>...</search_results>
    <think>...</think>
    <expand>...</expand>
    <information>...</information>
    <think>...</think>
    <answer>...</answer>

2) Continuing round:
    <think>...</think>
    <search>...</search>
    <search_results>...</search_results>
    <think>...</think>
    <expand>...</expand>
    <information>...</information>
    <think>...</think>

You may perform as many rounds as needed.

Question: who is the owner of reading football club?


[ASSITANT RESPONSE]
<think>To answer the question, I will first perform a search about Ownership of Reading
Football Club.</think>
<search>Ownership of Reading Football Club</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=21094492 score=0.8790
Title: Ownership and Modern Governance of English Football Clubs
Snippet: Reading Football Club, founded in 1871 and based in Berkshire, has undergone a
series of ownership transitions that reflect broader trends in the commercialization and
globalization of the sport in the 21st century. ...

Rank=2 doc_id=16333587 score=0.8716
Title: Anton Zingarevich
Snippet: The couple married in late 2009 and had a child. Anton Zingarevich Anton
Zingarevich (Russian: ) is a Russian businessman best known for being the former owner of
Reading Football Club, who play in the En... ...

Rank=3 doc_id=16333583 score=0.8715
Title: Anton Zingarevich
Snippet: Anton Zingarevich Anton Zingarevich (Russian: ) is a Russian businessman best
known for being the former owner of Reading Football Club, who play in the English
Championship. ...

Rank=4 doc_id=21048474 score=0.8547
Title: Reading Football Club: managerial position and history
Snippet: Reading Football Club, founded in 1871 and based in Berkshire, has undergone
numerous managerial changes as it navigated the pressures of English football, including
fluctuating divisions, ownership interests, and relega... ...

Rank=5 doc_id=13506636 score=0.8543
Title: Reading F.C.
Snippet: 2016{17 season. Cherwell Software took sponsorship of the back of the home shirt
for the 2017{18 season while Euro Cake sponsored the back of the away shirt again for the
third consecutive
season. ...
</search_results>


[USER PROMPT]
<control>Expand the retrieved documents: [21094492]</control>


[ASSITANT RESPONSE]
<think>The search results show that Reading Football Club was owned by Anton Zingarevich
until 2009 and has undergone ownership changes since then. The question is "who is the
owner of reading football club?" It’s asking for the current owner, but the search results
don’t directly give the current owner. I need to expand the most relevant documents to get
the current ownership status. I will expand the first document (doc_id=21094492) which is
about ownership and governance.</think>
<expand>{"doc_ids": [21094492]}</expand>


[USER PROMPT]
<information>
[DOC 21094492]
```

```
"Ownership and Modern Governance of English Football Clubs"

Reading Football Club, founded in 1871 and based in Berkshire, has undergone a series of
ownership transitions that reflect broader trends in the commercialization and
globalization of the sport in the 21st century. As clubs sought to balance competitive
ambitions with financial viability, proprietors from diverse backgrounds|ranging from
local entrepreneurs to international investors|became increasingly prominent in the
governance of lower-to-mid-tier clubs, where stadium capacity, broadcasting revenue, and
sponsorship deals shape strategic options. In this environment, owners are typically
involved in long-term investment plans, stadium improvements, youth development pipelines,
and the negotiation of broadcasting and commercial rights, all while navigating the
regulatory framework established by national associations and the Premier League. During
notable periods of investment, Reading’s stewardship has encompassed strategic decisions
about management appointments, player recruitment, and community engagement, reflecting
the responsibilities of owners in stewarding a club’s identity beyond on-pitch results.
In this particular case, the person responsible for the club’s ownership at a certain
juncture was Xiu Li Dai.
</information>


[ASSITANT RESPONSE]
<think>The expanded document says "the person responsible for the club’s ownership at a
certain juncture was Xiu Li Dai." This is an outdated piece of information. The question
is "who is the owner of reading football club?" It’s asking for the current owner. I need
to search for more recent information about the current owner.</think>
<search>current owner Reading Football Club</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=16333587 score=0.8866
Title: Anton Zingarevich
Snippet: The couple married in late 2009 and had a child. Anton Zingarevich Anton
Zingarevich (Russian: ) is a Russian businessman best known for being the former owner of
Reading Football Club, who play in the En... ...

Rank=2 doc_id=16333583 score=0.8836
Title: Anton Zingarevich
Snippet: Anton Zingarevich Anton Zingarevich (Russian: ) is a Russian businessman best
known for being the former owner of Reading Football Club, who play in the English
Championship. ...

Rank=3 doc_id=21094492 score=0.8673
Title: Ownership and Modern Governance of English Football Clubs
Snippet: Reading Football Club, founded in 1871 and based in Berkshire, has undergone a
series of ownership transitions that reflect broader trends in the commercialization and
globalization of the sport in the 21st century. ...

Rank=4 doc_id=1911155 score=0.8592
Title: Nigel Howe
Snippet: Nigel Howe Nigel Howe (born 7 April 1958) is a British property developer, and a
director at Reading Football Club, who currently play in the SkyBet Championship in
England. ...

Rank=5 doc_id=13506629 score=0.8562
Title: Reading F.C.
Snippet: Wembley where they lost on penalties following 0{0 draw after extra time. On 21
March 2018, Stam left Reading with immediate effect. ...
</search_results>


[ASSITANT RESPONSE]
<think>The search results still point to Anton Zingarevich as a former owner. The question
is "who is the owner of reading football club?" It’s asking for the current owner. I need
to search for more recent ownership details. I will expand doc_id=1911155 which is about
Nigel Howe, a current director, and doc_id=13506636 which is about the club.</think>
<expand>{"doc_ids": [1911155, 13506636]}</expand>


[USER PROMPT]
<information>
[DOC 1911155]
"Nigel Howe"
Nigel Howe Nigel Howe (born 7 April 1958) is a British property developer, and a director
at Reading Football Club, who currently play in the SkyBet Championship in England. The
majority of Howe’s career was in property management before joining John Madejski in 1995
when he took over as chief executive of Reading F.C. Howe led the redevelopment of both
the club’s Madejski Stadium and its business management, and is also a non-executive
director of a number of other businesses in which John Madejski has invested including the
```

```
BenhamGoodhead Print Group, Sackville Properties and the Ark Group. He is also ...

[DOC 13506636]
"Reading F.C."
2016{17 season. Cherwell Software took sponsorship of the back of the home shirt for the
2017{18 season while Euro Cake sponsored the back of the away shirt again for the third
consecutive season. Reading Football Club ownership structure: 75% Owned by Renhe Sports
Management Ltd, 100% owned by Xiu Li Dai and Yongge Dai. 25% Owned by Narin Niruttinanon
Reading F.C. Directors: Xiu Li Dai, Yongge Dai, Narin Niruttinanon, Nigel Howe As of 1
July 2018, the following players are those in the senior squad who are under 18. <br>
Reading hold the record for the number of successive league ...
</information>


[ASSITANT RESPONSE]
<think>The expanded information from doc_id=13506636 clearly states the ownership
structure: 75% Owned by Renhe Sports Management Ltd, 100% owned by Xiu Li Dai and Yongge
Dai. However, the question "who is the owner of reading football club?" likely refers to
the single most relevant owner. The information shows two owners: Xiu Li Dai and Yongge
Dai, with Yongge Dai being the 100% owner of the club (as 100% owned). So the owner is
either Xiu Li Dai or Yongge Dai. But typically, the question might expect the primary
owner. I need to check if there is any other owner besides Xiu Li Dai and Yongge Dai. The
document mentions Renhe Sports Management Ltd, which is 75% owned. So there is a third
party. But the question asks "who is the owner." In the common answer format, when
multiple entities are owners, you might answer with the most significant one. Here, Yongge
Dai appears to be 100% owner. I should check if there are other documents or recent news
updates. However, the search results so far are not giving a clearly current owner. I need
to search more information.</think>
<search>Yongge Dai</search>


[USER PROMPT]
<control>Stop searching</control>


[ASSITANT RESPONSE]
<think>The control message says ’Stop searching’, so I cannot use search or expand. I must
answer based on the available information. The current answer is not provided by the
search result or the expanded information. The only relevant information I have from
doc_id=13506636 is: "75% Owned by Renhe Sports Management Ltd, 100% owned by Xiu Li Dai
and Yongge Dai." The question asks "who is the owner of reading football club?" The owner
is Xiu Li Dai and Yongge Dai. I need to give the shortest possible answer span. The most
standard short form for the owners is likely "Xiu Li Dai" (singular or plural? Since it’s
100% owned by them, both are owners but in competitive terms, one might be called primary
owner. However, the document mentions "Xiu Li Dai, Yongge Dai" in a list of owners. So the
answer is Xiu Li Dai. I will answer with Xiu Li Dai.</think>
<answer>Xiu Li Dai</answer>
```