SCALING SEARCH-AUGMENTED LLM REASONING VIA ADAPTIVE INFORMATION CONTROL

Siheng Xiong, Oguzhan Gungordu, Blair Johnson, James C. Kerce, Faramarz Fekri
Georgia Institute of Technology
{sxiong45,ogungordu3}@gatech.edu, Blair.Johnson@gtri.gatech.edu,
Clayton.Kerce@gtri.gatech.edu, faramarz.fekri@ece.gatech.edu

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 only implicitly regulate agent behavior and provide limited guidance for deciding 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. Based on this utility, we introduce retrieval continuation and granularity control mechanisms that determine whether retrieval should proceed and which parts of hierarchical information to expand. These control signals provide explicit guidance during training, while an annealed control strategy enables the agent to gradually internalize effective information acquisition behaviors. Extensive experiments across seven benchmarks show that our method consistently outperforms strong outcome-based RL baselines and retrieval-based reasoning methods without explicit information control. 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].

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., 2025a).

To mitigate these issues, prior work (Jin et al., 2025a; 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. Agents often exhibit suboptimal retrieval behaviors: they may over-retrieve when evidence is weak, accumulating unnecessarily long contexts, or terminate retrieval prematurely even when additional evidence remains beneficial. These issues are particularly severe in long-horizon

[1]The code and data are available at https://github.com/xiongsiheng/DeepControl.

reasoning settings, where sparse outcome rewards provide little guidance for intermediate retrieval decisions. 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.

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 introduces explicit control signals that guide information acquisition by regulating retrieval granularity, deciding when to expand additional evidence, and determining when to halt 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. 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 $\mathcal{J}_{\text{PPO}}(\theta)$:

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

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 (see the failure cases in Section F):

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., 2025a) 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 (Xiong et al., b; 2025a). 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

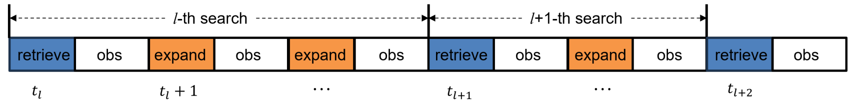


Figure 1: Definition of a search step. A search step starts with a retrieval action and includes all subsequent expansion actions until the next retrieval or termination.

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

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

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.

The proposed information utility provides dense, state-dependent feedback for intermediate retrieval decisions and exhibits diminishing marginal gains. In our implementation, effectiveness uses gold answers and is only used for training-time control; the final policy operates autonomously at test time. Detailed definitions and theoretical properties are provided in Section A.

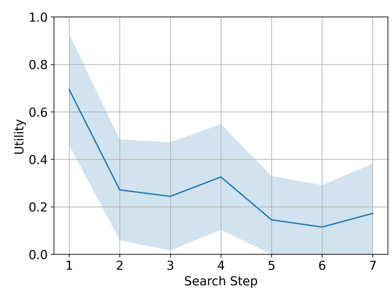


Figure 2: Information utility varies across search steps and does not always increase with additional retrieval, motivating explicit continuation control.

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

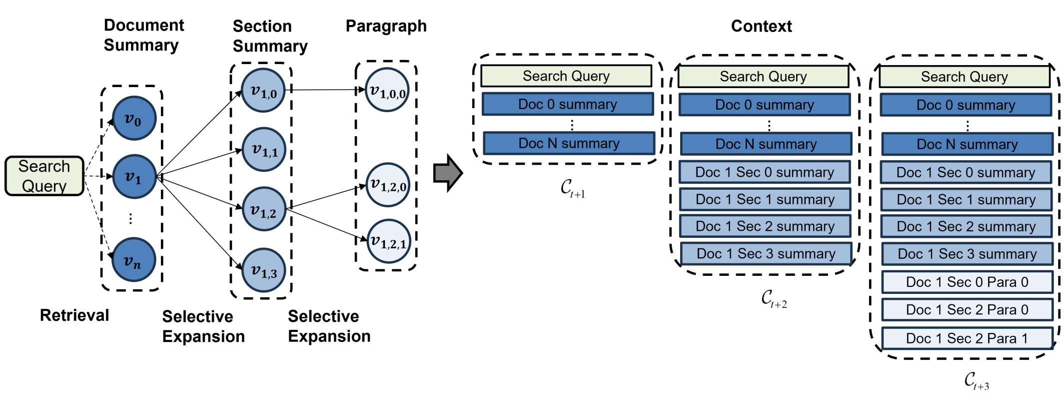


Figure 3: Hierarchical granularity control via selective expansion, where the agent incrementally refines retrieved information from coarse summaries to finer-grained content as needed.

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* triggers an additional retrieval step when recent utility remains high but the agent still lacks sufficient confidence to 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.

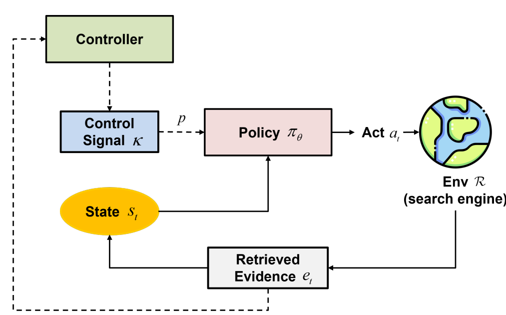


Figure 4: Trajectories generated in rollout mode with and without information control.

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.

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. †/⋆ 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., 2025a). For R1, rejection sampling and Search-R1, we use the fine-tuned version from (Jin et al., 2025a). Across all methods, we use the same retriever, corpus, effective retrieval budget, training data, and pretrained models.

Implementation details. We conduct experiments with Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct (Yang et al., 2024a). 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. (2025a), 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 E.

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 F. 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.

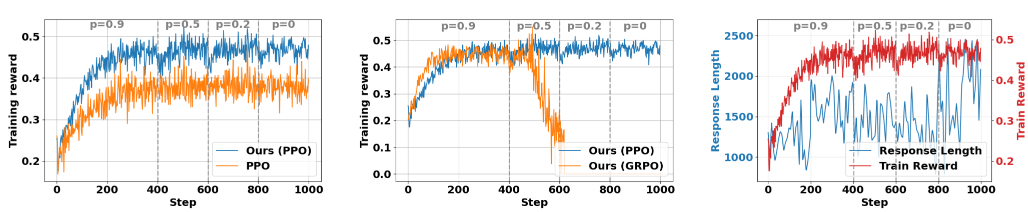
4.3 ANALYSIS

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 reaches higher reward throughout training under the same optimization setup. (b) GRPO leads to reward collapse, while PPO shows steadier optimization and maintains stable performance. (c) 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) demonstrate strong reasoning and coding abilities but often suffer from limited factual coverage and hallucinations (Zhang et al., 2023). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) addresses this issue by incorporating external documents 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., 2025a), which interleave reasoning steps with search calls. However, existing retrieval-augmented approaches largely assume that acquiring more information is beneficial and typically append retrieved content to the context using fixed or heuristic strategies. This often leads to redundant evidence accumulation, context saturation, and noisy reasoning in large information spaces. In contrast, our work explicitly regulates 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 used to optimize LLMs for complex behaviors such as reasoning and tool use. RLHF (Ouyang et al., 2022) and related methods (Rafailov et al., 2023) rely on preference-based rewards, while recent studies show that outcome-based RL can enable strong reasoning capabilities using only task-level supervision (Shao et al., 2024; Guo et al., 2025). Recent work on LLM reasoning has further explored structured exploration strategies and process-level supervision/verification to improve generation diversity and training/inference stability (Yang et al., 2024c;b; Xiong et al., 2024; 2025b; a; Yu et al., 2025; Gungordu et al., 2026). Several works extend RL to tool-augmented agents, enabling models to learn when and how to invoke external tools (Nakano et al., 2021; Jin et al., 2025a). However, most approaches rely primarily on sparse outcome-level rewards, which provide limited guidance for intermediate decisions such as whether to continue retrieval or how much information to acquire. Consequently, agents often exhibit inefficient retrieval behaviors, including over-retrieval and premature termination. Our work addresses this limitation by introducing explicit information control signals derived from information utility to guide retrieval behaviors during training.

Information Control in Search-Augmented Reasoning. Effective exploration remains a central challenge in reinforcement learning. Prior work proposes intrinsic rewards, count-based exploration, and curiosity-driven objectives to encourage novelty and state coverage (Pathak et al., 2017; Bellemare et al., 2016). Related efforts in sequential decision-making also study adaptive computation and action selection under uncertainty, highlighting the importance of allocating limited resources to informative decisions (Russell et al., 1991; Zilberstein, 1996; 2011). However, these approaches do not explicitly address information acquisition in search-augmented reasoning. In particular, they do not model the utility of retrieved information nor provide mechanisms to regulate retrieval continuation or granularity. Our work fills this gap by introducing a utility-driven framework that treats information acquisition as a controllable process and enables agents to internalize effective retrieval behaviors through annealed control.

6 CONCLUSION

We introduce an adaptive information control framework that regulates information acquisition in search-augmented reasoning agents using information utility. 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 that allows the model to gradually internalize appropriate information acquisition behaviors without external intervention. Experiments across multiple tasks and model scales demonstrate improvements in reasoning accuracy, training stability, and computational efficiency, highlighting the importance of treating information acquisition as a controllable and learnable process. Future work may explore richer definitions of information utility, uncertainty-aware retrieval and stopping criteria, and extensions to multi-tool and multimodal reasoning settings.

ACKNOWLEDGMENTS

This work is supported in part by DARPA SciFy program, Award No.HR001125C0302.

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.

Oguzhan Gungordu, Siheng Xiong, and Faramarz Fekri. Pathwise: Planning through world model for automated heuristic design via self-evolving llms. *arXiv preprint arXiv:2601.20539*, 2026.

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*, 2025a.

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (eds.), *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pp. 737–740. ACM, 2025b. doi: 10.1145/3701716.3715313. URL `https://doi.org/10.1145/3701716.3715313`.

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.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

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.

Stuart Russell, Eric H Wefald, Daniel G Bobrow, Michael Brady, and Randall Davis. Do the right thing. *(No Title)*, 1991.

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, Zhangding Liu, Jieyu Zhou, and Yusen Su. Deliberate planning in language models with symbolic representation. In *Twelfth Annual Conference on Advances in Cognitive Systems*, a.

Siheng Xiong, Ali Payani, and Faramarz Fekri. Enhancing language model reasoning with structured multi-level modeling. In *The Fourteenth International Conference on Learning Representations*, b.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10452–10470, 2024.

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*, 2024a.

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. Can llms reason in the wild with programs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9806–9829, 2024b.

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. Harnessing the power of large language models for natural language to first-order logic translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6942–6959, 2024c.

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.

Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. Causaleval: Towards better causal reasoning in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12512–12540, 2025.

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.

Shlomo Zilberstein. Using anytime algorithms in intelligent systems. *AI magazine*, 17(3):73–73, 1996.

Shlomo Zilberstein. Metareasoning and bounded rationality., 2011.

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 $t = 0, 1, \dots, T-1$ index primitive actions (e.g., `retrieve`, `expand`, `answer`), and let $l = 0, 1, \dots, L-1$ 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

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

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

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

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

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

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 *helpful* for solving the task, i.e., whether it increases the model's likelihood of a correct answer. Unlike novelty, effectiveness is computed with respect to the net injected nodes contributed by the search step, $\Delta\mathcal{C}_l$, which may include both internal and leaf nodes.

Let $\mathcal{Y}^*(u)$ denote the set of acceptable gold answer strings (aliases) for task u. To isolate the effect of injected evidence from stochastic variations in reasoning, we condition the language model on the task u, the injected evidence, and a fixed reasoning trace c, where c is generated via deterministic decoding under each evidence condition. Concretely, let $\mathcal{C}_{t_{l+1}-1}$ denote the injected evidence accumulated up to the end of the l-th search step. For each target string $y \in \mathcal{Y}^*(u)$, we compute a length-normalized mean log-likelihood:

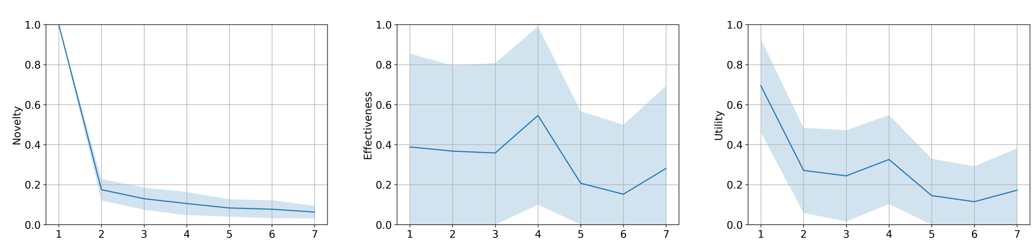
$$s_l(y) = \frac{1}{|y|} \sum_{i=1}^{|y|} \log \mathbb{P}\big(y_i \mid y_{<i},\, u,\, \mathcal{C}_{t_{l+1}-1},\, c\big)\,, \tag{10}$$

where $|y|$ is the number of tokens in y. We aggregate across aliases using log-mean-exp:

$$S_l = \log\left(\frac{1}{|\mathcal{Y}^*(u)|} \sum_{y \in \mathcal{Y}^*(u)} \exp\big(s_l(y)\big)\right). \tag{11}$$

Effectiveness is defined as the *positive improvement* in this target score induced by the newly injected evidence of the l-th search step:

$$\Delta_l = \max\big(0,\ S_l - S_{l-1}\big). \tag{12}$$



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

Figure 6: Information novelty, effectiveness, and utility as functions of the search step. Lines denote the mean across rollouts and shaded regions represent one standard deviation.

To obtain a bounded score, we rescale Δ_l to $[0, 1]$ using two thresholds $\tau_{\text{low}} < \tau_{\text{high}}$:

$$\text{Effectiveness}(\Delta\mathcal{C}_l \mid u, s_{t_l}) = \begin{cases} 0, & \Delta_l \leq \tau_{\text{low}}, \\ \frac{\Delta_l - \tau_{\text{low}}}{\tau_{\text{high}} - \tau_{\text{low}}}, & \tau_{\text{low}} < \Delta_l < \tau_{\text{high}}, \\ 1, & \Delta_l \geq \tau_{\text{high}}. \end{cases} \tag{13}$$

By construction, effectiveness is high only when newly injected evidence increases the model's confidence on the gold answer, and is zero when the evidence decreases or does not improve it. Note that this effectiveness signal is used only during *training*, when gold answers are available.

We illustrate how novelty, effectiveness, and utility evolve with additional evidence in Figure 6 (see Section E for hyperparameters used in our paper). While novelty rapidly decreases after the first retrieval step, effectiveness remains non-zero in later steps, suggesting that later evidence is often less novel but still useful for improving answer confidence.

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

1) **Monotonicity with novel and beneficial evidence.** When newly retrieved evidence is both novel with respect to the current reasoning state and increases the model's confidence on the gold answer (i.e., yields positive effectiveness), the information utility increases accordingly. Conversely, evidence that is redundant or does not improve the gold-answer likelihood yields little utility gain.
2) **Diminishing returns after task completion.** After sufficient evidence for solving the task has been acquired, additional retrievals tend to be increasingly redundant and provide only limited improvement to the gold-answer likelihood, leading to diminishing marginal utility.

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

Discussion. We use information utility as an **external control signal**, rather than incorporating it directly into the RL reward. This distinguishes *explicit* regulation of information acquisition (via control messages that can intervene at specific steps) from *implicit* learning of such behaviors through reward shaping. This design choice is motivated by: 1) separating utility estimation from policy optimization makes the framework modular, allowing the controller and utility definition to be iterated or replaced without changing the underlying RL objective or training pipeline; 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) **Target-score saturation.** After sufficient task-relevant evidence has been incorporated into the context, the model's confidence on the gold answer saturates, i.e., the aggregated target score S_l (defined in Section A) becomes approximately stable:

$$S_l \approx S_{l-1}, \tag{14}$$

for sufficiently large l.

(A2) **Evidence redundancy accumulation.** As the number of search steps increases, newly retrieved evidence becomes 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 yields positive effectiveness, then the information utility is strictly positive.*

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

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

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

$$U(e_l \mid u, s_{t_l}) = \rho \cdot \mathrm{Novelty}(e_l \mid s_{t_l}) + (1 - \rho) \cdot \mathrm{Effectiveness}(e_l \mid u, s_{t_l}) > 0. \tag{15}$$

□

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 beneficial for the downstream task (as reflected by positive effectiveness).

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 aggregated target score S_l saturates after sufficient task-relevant evidence has been incorporated. Recall that our effectiveness is defined from the positive improvement $\Delta_l = \max(0, S_l - S_{l-1})$, optionally rescaled to $[0, 1]$ via fixed thresholds. Therefore,

$$\lim_{l \to \infty} \Delta_l = 0 \qquad \Rightarrow \qquad \lim_{l \to \infty} \mathrm{Effectiveness}(e_l \mid u, s_{t_l}) = 0. \tag{16}$$

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

$$\lim_{l \to \infty} \mathrm{Novelty}(e_l \mid s_{t_l}) = 0. \tag{17}$$

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

$$\lim_{l \to \infty} U(e_l \mid u, s_{t_l}) = 0. \tag{18}$$

□

Lemma 2 captures diminishing returns in an asymptotic sense: after sufficient task-relevant information has been acquired, additional retrievals tend to introduce increasingly redundant evidence and provide limited improvement to the gold-answer likelihood, 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

$$\mathcal{C}^*_{t'} = \mathcal{C}^*_{t'-1} \cup \{ v' \mid v \in \mathcal{C}^*_{t'-1}, (v, v') \in \xi^*_{t'} \}. \tag{19}$$

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_{stop} for m_{stop} consecutive search steps, we define the stopping index

$$l^\star = \min_{l \in [m_{\text{stop}}-1, L-1]} \max_{j \in [l-m_{\text{stop}}+1, l]} U(e_j) < \delta_{\text{stop}}. \tag{20}$$

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. We trigger a one-shot continuation intervention when (i) the utility of the most recent m_{cont} search steps remains consistently high ($\geq \delta_{\text{cont}}$), but (ii) the model is still insufficiently confident on the gold answer under the current evidence. Concretely, let S_l denote the aggregated target score (defined in Section A) computed under evidence $\mathcal{C}_{t_{l+1}-1}$. If the agent attempts to terminate at search-step index l and

$$S_l \leq \tau_{\text{score}} \ \wedge \min_{j \in [l-m_{\text{cont}}+1, l]} U(e_j) \geq \delta_{\text{cont}}, \tag{21}$$

we inject a one-shot control signal κ = `Continue the search for one additional step`. Here τ_{score} is a confidence threshold on the gold-answer score. Note that Equation (21) is used only during *training* when gold answers are 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 E.

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 E.

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

$$\hat{r}_\phi(\tau, y_{\text{gold}}) \;=\; r_{\text{correct}}(\tau, y_{\text{gold}}) + r_{\text{penalty}}(\tau) + r_{\text{ret}}(\tau, y_{\text{gold}}). \tag{22}$$

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

The base reward of correctness is

$$r_{\text{correct}}(\tau, y_{\text{gold}}) \;=\; \begin{cases} \max\big(\mathrm{F1}(y_{\text{pred}}, y_{\text{gold}}),\ \lambda_{\text{format}}\big), & y_{\text{pred}} \neq \varnothing, \\ 0, & \text{otherwise}, \end{cases} \tag{23}$$

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

$$r_{\text{penalty}}(\tau) \;=\; -\min\big(\lambda_{\text{penalty}} \cdot N_{\text{penalty}}(\tau),\ \lambda_{\text{penalty}}^{\max}\big), \tag{24}$$

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

$$r_{\text{ret}}(\tau, y_{\text{gold}}) \;=\; \begin{cases} \lambda_{\text{ret}}, & \mathbb{I}_{\text{ret}}(\tau, y_{\text{gold}}) = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{25}$$

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.

$$r_{\phi}(\tau, y_{\text{gold}}) \;=\; \begin{cases} \min(\hat{r}_{\phi}(\tau, y_{\text{gold}}), 1), & \text{F1}(y_{\text{pred}}, y_{\text{gold}}) = 1, \\ \min(\hat{r}_{\phi}(\tau, y_{\text{gold}}), \lambda_{\text{ceil}}), & \text{otherwise}, \end{cases} \tag{26}$$

Detailed hyperparameter settings and ablations are provided in Section E.

D Dataset Overview

We evaluate DeepControl on two categories of tasks: general question answering and multi-hop question answering. For general question answering, we use Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2022). For multi-hop question answering, we evaluate on HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), Musique (Trivedi et al., 2022b), and Bamboogle (Press et al., 2022). All dataset splits are obtained from the FlashRAG toolkit (Jin et al., 2025b) via its curated dataset collection.

Natural Questions consists of real Google search queries paired with Wikipedia answers annotated by humans (79,168 training and 3,610 test samples). TriviaQA is a large-scale reading comprehension benchmark; we use its 11,313-example test set. PopQA contains 14,267 entity-centric triples designed to measure parametric knowledge coverage on long-tail entities. HotpotQA is a crowd-sourced Wikipedia-based multi-hop dataset requiring reasoning across multiple paragraphs (90,447 training and 7,405 development samples). 2WikiMultiHopQA combines structured and unstructured Wikipedia information; we evaluate on its 12,576-example development split. Musique composes single-hop questions into 2–4 hop problems; we use its 2,417-example development set. Bamboogle is a manually curated set of 125 two-hop compositional questions selected because search engines originally answered them incorrectly. Examples from each dataset are provided in Section F.

E Implementation Details

Prompts. In Section F, we present all prompts used in our framework, including the search-augmented reasoning prompt and the control message.

Control hyperparameters. Unless otherwise specified, we use a unified set of control hyperparameters across all tasks (see definitions in Section A). We embed retrieved passages using the E5 encoder (`intfloat/e5-base-v2`) (Wang et al., 2022), truncate each passage to at most 512 tokens for encoding, and compute novelty via a k-NN estimator with $k_{\text{nn}} = 5$. Effectiveness is computed from the positive improvement in the aggregated gold-answer score S_l, using $\Delta_l = \max(0, S_l - S_{l-1})$. We rescale Δ_l to $[0, 1]$ with thresholds $\tau_{\text{low}} = 0.3$ and $\tau_{\text{high}} = 3.0$. The target score S_l is computed under a deterministic reasoning trace with a maximum of 128 CoT tokens. We combine novelty and effectiveness as utility, with $\rho = 0.5$. We stop searching when utility stays below a threshold $\delta_{\text{stop}} = 0.2$ for $m_{\text{stop}} = 2$ consecutive search steps. We consider a one-shot continuation intervention when the recent utility remains high but the model is still insufficiently confident on the gold answer. Concretely, we require the utility to exceed a high-utility threshold $\delta_{\text{cont}} = 0.3$ for $m_{\text{cont}} = 2$ consecutive search steps, and additionally require the current gold-answer score S_l to be below a confidence threshold $\tau_{\text{score}} = -2.0$ (note that S_l is a length-normalized log-likelihood score and is typically negative). We observe that the overall control behavior is insensitive to moderate 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., 2024a). 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., 2025a), 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 F. All experiments are conducted with a fixed random seed.

F ADDITIONAL RESULTS

DEEPCONTROL vs. Vanilla 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 under information control. 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. In Section F, we present representative examples of DEEPCONTROL under settings with and without control signals, evaluated on both single-hop and multi-hop questions. These examples illustrate how the agent interleaves reasoning with retrieval and selective expansion, while control messages guide the agent to regulate search behavior. In particular, the examples highlight how continuation and termination controls help avoid unnecessary retrieval steps and improve decision efficiency during reasoning.

Example: Search-Augmented Reasoning Tasks

```
[NQ]
Question: who has the right of way in international waters
Answer: Neither vessel

Question: how many episodes are in series 7 game of thrones
Answer: seven


[TriviaQA]
Question: Who wrote the novel Evening Class?
Answer: Maeve Binchy

Question: Which item of clothing cost Isadora Duncan her life?
Answer: Scarf- it was caught in a moving car wheel


[PopQA]
Question: What is the capital of Mali?
Answer: Bamako

Question: What is Dominick Bellizzi's occupation?
Answer: jockey


[HotpotQA]
Question: Dixit Dominus is a psalm setting by who, which was a German, later British,
baroque composer who spent the bulk of his career in London, becoming well known for his
operas, oratorios, anthems, and organ concertos?
Answer: George Frideric Handel

Question: Zimbabwe's Guwe Secondary School has a sister school in what New York county?
Answer: Nassau County


[2WikiMultiHopQA]
Question: When did John V, Prince Of Anhalt-Zerbst's father die?
Answer: 12 June 1516

Question: Who is older, Aryeh Ben-Eliezer or Jason Pociask?
Answer: Aryeh Ben-Eliezer


[Musique]
Question: Where is Ulrich Walter's employer headquartered?
Answer: Cologne

Question: Who founded the company that distributed the film UHF?
Answer: Mike Medavoy


[Bamboogle]
Question: The main actor of Indiana Jones is a licensed what?
Answer: pilot

Question: How many people died in the second most powerful earthquake ever recorded?
Answer: 131
```

Example: Suboptimal Search Behavior (Search-R1-qwen2.5-3b-it-grpo on NQ)

```
[USER PROMPT]
Answer the given question. You must conduct reasoning inside <think> and </think> first
      every time you get new information. After reasoning, if you find you lack some
     knowledge, you can call a search engine by <search> query </search> and it will
     return the top searched results between <information> and </information>. You can
     search as many times as your want. If you find no further external knowledge
     needed, you can directly provide the answer inside <answer> and </answer>, without
      detailed illustrations. For example, <answer> Beijing </answer>.
Question: when is the next scandal episode coming out?

[ASSISTANT RESPONSE]
To determine when the next Scandal episode is coming out, I will follow these steps:

1. Identify the current season of Scandal.
2. Find out the date of the last episode of the current season.
```

```
3. Determine the date of the next episode of the current season.
4. Provide the date of the next episode if available.

Let's start with the first step.

<search>Scandal current season</search>
<information>
Doc 1(Title: "Scandal (TV series)") ABC renewed ""Scandal"" for a sixth season, which
     premiered on January 26, 2017. The entire season is about the murder of President-
     elect Frankie Vargas. It is revealed that his wife, Luna, was behind his murder
     because she wanted power. The end of season six concludes with Mellie becoming the
      first female President of the United States. At the end of Fitz's term, he
     decided to reinstate a B613 and become Command, due to Rowan's manipulation of him
     . Olivia convinces him not to, because she thinks it will destroy him. Then she
     has Mellie reserve a percentage of the Pentagon's funds
Doc 2(Title: "Scandal (TV series)") in a blind emergency trust and uses that money to
     reinstate B613 herself. Olivia becomes Command at the end of the season and
     considers herself the most powerful person in the world. On February 10, 2017, ABC
      renewed ""Scandal"" for a seventh season; on May 10, 2017, it was announced that
     this season would be the show's last. On January 3, 2018, a crossover with ""How
     to Get Away with Murder"" was announced, which aired on March 1, 2018, with Viola
     Davis and Aja Naomi King guest starring. In early 2011, it was announced that
     Shonda Rhimes was developing a
Doc 3(Title: "Scandal (TV series)") Scandal (TV series) Scandal is an American
     political thriller television series starring Kerry Washington. Created by Shonda
     Rhimes, it aired on ABC from April 5, 2012, until April 19, 2018, for 124 episodes
      over seven seasons. Kerry Washington's character, Olivia Pope, is partially based
      on former George H.W. Bush administration press aide Judy Smith, who serves as a
     co-executive producer. The show takes place in Washington, D.C. and focuses on
     Olivia Pope's crisis management firm, Olivia Pope & Associates (OPA), and its
     staff, as well as staff at the White House and surrounding political scene. In
     addition to Kerry Washington,
</information>

From the search results, I can see that Scandal is currently in its seventh season. The
      last episode of the current season premiered on January 26, 2017.

To find the date of the next episode, I will search again.

<search>Scandal next episode date</search>
<information>
Doc 1(Title: "Scandal (TV series)") in a blind emergency trust and uses that money to
     reinstate B613 herself. Olivia becomes Command at the end of the season and
     considers herself the most powerful person in the world. On February 10, 2017, ABC
      renewed ""Scandal"" for a seventh season; on May 10, 2017, it was announced that
     this season would be the show's last. On January 3, 2018, a crossover with ""How
     to Get Away with Murder"" was announced, which aired on March 1, 2018, with Viola
     Davis and Aja Naomi King guest starring. In early 2011, it was announced that
     Shonda Rhimes was developing a
Doc 2(Title: "Scandal (season 6)") Scandal (season 6) The sixth season of the American
     television drama series ""Scandal"" was ordered on March 3, 2016 by ABC. It began
     airing on January 26, 2017 in the United States on ABC. The season includes the
     100th episode of the series, being the tenth episode of this season. The season
     was produced by ABC Studios, in association with ShondaLand Production Company;
     the showrunner being Shonda Rhimes. The season continues the story of Olivia Pope'
     s crisis management firm, Olivia Pope & Associates, and its employees, as well as
     staff at the White House in Washington, D.C. Season six had
Doc 3(Title: "Scandal (season 6)") character was killed off in the eleventh episode.
     The review aggregator website Rotten Tomatoes reports an 80% approval rating with
     an average rating of 7.5/10 based on 10 reviews. The website's consensus reads, ""
     As shocking and thrilling as ever, ""Scandal"" eases into its sixth season with as
      much confidence as Olivia Pope herself."" Scandal (season 6) The sixth season of
     the American television drama series ""Scandal"" was ordered on March 3, 2016 by
     ABC. It began airing on January 26, 2017 in the United States on ABC. The season
     includes the 100th episode of the series, being the tenth episode
</information>

From the search results, I can see that the next episode of the current season of
     Scandal premiered on January 26, 2017.

<answer>January 26, 2017</answer>

Therefore, the next episode of Scandal is coming out on January 26, 2017.
```

Example: Suboptimal Search Behavior (Search-R1-qwen2.5-3b-it-grpo on HotpotQA)

```
[USER PROMPT]
Answer the given question. You must conduct reasoning inside <think> and </think> first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine by <search> query </search> and it will return the top searched results between <information> and </information>. You can search as many times as your want. If you find no further external knowledge needed, you can directly provide the answer inside <answer> and </answer>, without detailed illustrations. For example, <answer> Beijing </answer>.
Question: The football manager who recruited David Beckham managed Manchester United during what timeframe?

[ASSISTANT RESPONSE]
To determine the timeframe during which the football manager who recruited David Beckham managed Manchester United, I will follow these steps:

1. Identify the football manager who recruited David Beckham.
2. Find the timeframe during which this manager managed Manchester United.

Let's start with step 1: identifying the football manager who recruited David Beckham.

<search>David Beckham recruited football manager</search>
<information>
Doc 1(Title: "David Beckham") David Beckham David Robert Joseph Beckham, (; born 2 May 1975) is an English retired professional footballer and current President of Inter Miami CF. He played for Manchester United, Preston North End, Real Madrid, Milan, LA Galaxy, Paris Saint-Germain and the England national team, for which he held the appearance record for an outfield player until 2016. He is the first English player to win league titles in four countries: England, Spain, the United States and France. He retired in May 2013 after a 20-year career, during which he won 19 major trophies. Known for his range of passing, crossing
Doc 2(Title: "David Beckham") Beckham's many endorsement deals make him one of the most recognisable athletes throughout the world. Having worn Adidas football boots from the start of his career (notably Adidas Predator), in 2003 he signed a $160 million lifetime contract with Adidas, earning nearly half the money upfront, and will continue to earn percentages of profits on all of his branded Adidas products . He had a 10-year collaboration with Pepsi Co. that expired in 2009. He has also promoted The Walt Disney Company theme parks. Beckham has several eponymous video games, including ""Go! Go! Beckham! Adventure on Soccer Island"", a platform game
Doc 3(Title: "Simon Fuller") Beckham ownership of their own apparel business specifically to develop the David Beckham brand and other iconic branded consumer products. The first result of this partnership was the Fall 2016 re-launch of the British heritage menswear label Kent & Curwen, which under Beckham, Fuller and 7 Global ownership currently runs 113 stores in 51 cities worldwide. Fuller first ventured into sports management in the mid-1990s, when he negotiated the historic first post Bosman transfer of England footballer Steve McManaman to Real Madrid. In 2003 Fuller formed a joint venture partnership with David Beckham, and later masterminded Beckham's $250 million move
</information>

From the search results, we can see that David Beckham was recruited by Simon Fuller. Now, let's move to step 2: finding the timeframe during which Simon Fuller managed Manchester United.

<search>Simon Fuller managed Manchester United</search>
<information>
Doc 1(Title: "Simon Fuller") Beckham ownership of their own apparel business specifically to develop the David Beckham brand and other iconic branded consumer products. The first result of this partnership was the Fall 2016 re-launch of the British heritage menswear label Kent & Curwen, which under Beckham, Fuller and 7 Global ownership currently runs 113 stores in 51 cities worldwide. Fuller first ventured into sports management in the mid-1990s, when he negotiated the historic first post Bosman transfer of England footballer Steve McManaman to Real Madrid. In 2003 Fuller formed a joint venture partnership with David Beckham, and later masterminded Beckham's $250 million move
Doc 2(Title: "Simon Fuller") to LA Galaxy. He remains Beckham's closest advisor. In 2009, Fuller took on personal representation of tennis player Andy Murray, guiding him to Grand Slam and Wimbledon wins and later helping the player to establish his own business. In March 2011, Simon Fuller was announced as the manager of British F1 driver Lewis Hamilton and surprised many when he moved Hamilton from McLaren and signed him to a long term contract with the MercedesAMG team. Hamilton won the 2014 F1 World Championship and publicly thanked Fuller, although Hamilton did not renew his deal with Fuller's XIX Entertainment company once it
Doc 3(Title: "Simon Fuller") expired at the end of that year. In 2013 it was announced that Fuller was managing 2012 Tour De France winner, Bradley Wiggins, helping him to set up his own pro-cycling team. In 2014, Fuller and Beckham discussed plan to buy an MLS football team in Miami and later the same year Fuller was revealed as
```

```
     the mastermind behind Beckham's involvement in a new global whisky launch with
     leading British drinks company Diageo. Fuller's longterm plan to deliver Beckham's
      ownership of an MLS team was announced in January 2018 when Miami was confirmed
     as 25th MLS team. Simon Fuller has
</information>

To determine the timeframe during which the football manager who recruited David
     Beckham managed Manchester United, I will follow these steps:

1. Identify the football manager who recruited David Beckham.
2. Find the timeframe during which this manager managed Manchester United.

From the previous search, we already know that the football manager who recruited David
      Beckham is Simon Fuller. Now, let's find the timeframe during which Simon Fuller
     managed Manchester United.

<search>Simon Fuller managed Manchester United timeframe</search>
<information>
Doc 1(Title: "Simon Fuller") Beckham ownership of their own apparel business
     specifically to develop the David Beckham brand and other iconic branded consumer
     products. The first result of this partnership was the Fall 2016 re-launch of the
     British heritage menswear label Kent & Curwen, which under Beckham, Fuller and 7
     Global ownership currently runs 113 stores in 51 cities worldwide. Fuller first
     ventured into sports management in the mid-1990s, when he negotiated the historic
     first post Bosman transfer of England footballer Steve McManaman to Real Madrid.
     In 2003 Fuller formed a joint venture partnership with David Beckham, and later
     masterminded Beckham's $250 million move
Doc 2(Title: "Simon Fuller") to LA Galaxy. He remains Beckham's closest advisor. In
     2009, Fuller took on personal representation of tennis player Andy Murray, guiding
      him to Grand Slam and Wimbledon wins and later helping the player to establish
     his own business. In March 2011, Simon Fuller was announced as the manager of
     British F1 driver Lewis Hamilton and surprised many when he moved Hamilton from
     McLaren and signed him to a long term contract with the MercedesAMG team. Hamilton
      won the 2014 F1 World Championship and publicly thanked Fuller, although Hamilton
      did not renew his deal with Fuller's XIX Entertainment company once it
Doc 3(Title: "Simon Fuller") expired at the end of that year. In 2013 it was announced
     that Fuller was managing 2012 Tour De France winner, Bradley Wiggins, helping him
     to set up his own pro-cycling team. In 2014, Fuller and Beckham discussed plan to
     buy an MLS football team in Miami and later the same year Fuller was revealed as
     the mastermind behind Beckham's involvement in a new global whisky launch with
     leading British drinks company Diageo. Fuller's longterm plan to deliver Beckham's
      ownership of an MLS team was announced in January 2018 when Miami was confirmed
     as 25th MLS team. Simon Fuller has
</information>

<answer>2003-2009</answer>
```

Example: Output Format Collapse under Large Top-K Retrieval (Search-R1-qwen2.5-3b-it-grpo on HotpotQA)

```
[USER PROMPT]
Answer the given question. You must conduct reasoning inside <think> and </think> first
every time you get new information. After reasoning, if you find you lack some knowledge,
you can call a search engine by <search> query </search> and it will return the top
searched results between <information> and </information>. You can search as many times as
your want. If you find no further external knowledge needed, you can directly provide the
answer inside <answer> and </answer>, without detailed illustrations. For example,
<answer> Beijing </answer>.
Question: What government position was held by the woman who portrayed Corliss Archer in
the film Kiss and Tell?

[ASSISTANT RESPONSE]
To determine which government position was held by the woman who portrayed Corliss Archer
in the film Kiss and Tell, I will follow these steps:

1. Identify the woman who portrayed Corliss Archer in the film Kiss and Tell.
2. Find out which government position she held.

Let's start with step 1: identifying the woman who portrayed Corliss Archer in the film
Kiss and Tell.

<search>Corliss Archer Kiss and Tell </search>
<information>
Doc 1 (Title: "Kiss and Tell (play)")
1945. The play was produced by George Abbott and written by F. Hugh Herbert. It ran for a
```

```
total of 956 performances before closing on June 23, 1945. Two teenage girls become
interested in boys, and the girls' parents are making more problems rather than solving
them. A film version written by Herbert and based on his play was released by Columbia
Pictures on October 4, 1945 with Shirley Temple in the role of Corliss Archer. The film
was directed by Richard Wallace and produced by Sol C. Siegel. Temple would reprise the
role in the film's sequel "A Kiss

Doc 2 (Title: "Kiss and Tell (1945 film)")
Kiss and Tell (1945 film) Kiss and Tell is a 1945 American comedy film starring then
17-year-old Shirley Temple as Corliss Archer. In the film, two teenage girls cause their
respective parents much concern when they start to become interested in boys. The parents'
bickering about which girl is the worse influence causes more problems than it solves. The
movie was based on the Broadway play "Kiss and Tell", which was based on the Corliss
Archer short stories. The stories, play and movie were all written by F. Hugh Herbert. A
sequel film, "A Kiss for Corliss", was released in

Doc 3 (Title: "Meet Corliss Archer")
The program was produced by Ziv Productions. Several episodes of the Ziv version are
available on DVD, and some are also available from the Internet Archive. Robin Morgan
portrayed Corliss in a live telecast of "Kiss and Tell" on "The Alcoa Hour" (August 5,
1956), with Warren Berlinger as Dexter. Radio listeners had to use their imaginations to
visualize Corliss, her friends and her town. But those imaginations got a boost in 1948
when the "Meet Corliss Archer" comic book, published by Fox Feature Syndicate, came out in
three issues from March to July 1948. Al Feldstein (Albert B. Feldstein),

Doc 4 (Title: "A Kiss for Corliss")
A Kiss for Corliss A Kiss for Corliss is a 1949 American comedy film directed by Richard
Wallace and written by Howard Dimsdale. It stars Shirley Temple in her final starring role
as well as her final film appearance. It is a sequel to the 1945 film "Kiss and Tell". "A
Kiss for Corliss" was retitled "Almost a Bride" before release and this title appears in
the title sequence. The film was released on November 25, 1949, by United Artists.
Corliss Archer (Shirley Temple) is an impulsive teenager excited by notorious playboy
Kenneth Marquis (David Niven), who has already been

Doc 5(Title: "Kiss and Tell (play)") Kiss and Tell (play) Kiss and Tell is a 1943
Broadway play by F. Hugh Herbert. ""Kiss and Tell"" starred Joan Caulfield as Corliss
Archer and Judith Parrish as her friend Mildred Pringle. The play's great success led to
offers from Hollywood for Caulfield, who left the production almost one year to the day
from when it opened, and was replaced in her role by her sister Betty Caulfield. ""Kiss
and Tell"" opened at the Biltmore Theatre in New York City on March 17, 1943. It remained
there until the end of 1944, before transferring to the Bijou Theatre in

Doc 6(Title: "Kiss and Tell (1945 film)") reported that the film hadn't changed much from
the ""protracted but reasonably diverting"" stage version, and singled out Jerome
Courtland as ""one of the most capable adolescent actors to come along in years."" Kiss
and Tell (1945 film) Kiss and Tell is a 1945 American comedy film starring then
17-year-old Shirley Temple as Corliss Archer. In the film, two teenage girls cause their
respective parents much concern when they start to become interested in boys. The parents'
bickering about which girl is the worse influence causes more problems than it solves. The
movie was based on the Broadway play ""Kiss

Doc 7(Title: "Kiss and Tell (play)") for Corliss"" in 1949. Kiss and Tell (play) Kiss and
Tell is a 1943 Broadway play by F. Hugh Herbert. ""Kiss and Tell"" starred Joan Caulfield
as Corliss Archer and Judith Parrish as her friend Mildred Pringle. The play's great
success led to offers from Hollywood for Caulfield, who left the production almost one
year to the day from when it opened, and was replaced in her role by her sister Betty
Caulfield. ""Kiss and Tell"" opened at the Biltmore Theatre in New York City on March 17,
1943. It remained there until the end of 1944, before transferring to

Doc 8(Title: "Meet Corliss Archer") of a series of stories. ""Kiss and Tell"" was a 1943
play that was adapted for a 1945 film starring Shirley Temple. The 1949 sequel, ""A Kiss
for Corliss"", was re-released in 1954. Like many other radio shows, ""Meet Corliss
Archer"" made the leap to television with live performances in 1951 and 1952, and from
1954 to 1955, as a syndicated television show starring Ann Baker and Mary Brian. One of
the show's unique features was the occasional cut to a comic-book-style drawing, with
announcer's commentary, that illustrated the current story situation and was used several
times during each episode.

Doc 9(Title: "Kiss and Tell (1945 film)") 1949 and also starred Temple, but was not
written by Herbert. To boost sales and attract customers at the local bazaar,
fifteen-year-old Corliss Archer and seventeen-year-old Mildred Pringle decide to start
selling kisses. When their booth at a USO bazaar fails to attract customers, teenager
Corliss Archer suggests to her best friend, Mildred Pringle, that they sell kisses. The
idea becomes a success among the soldiers visiting the bazaar, and business is booming,
until the girls' mothers find out about it. Despite the fact that it was Corliss' idea,
Mrs. Archer blames Mildred for the girls' behavior. Mrs. Archer's assumptions

Doc 10(Title: "Kiss and Tell (1945 film)") greatly upset Mr. Pringle when he hears about
them at the Pringle dinner table. He decides the family, including Mildred and their son
```

```
Raymond, will refrain from socializing with the Archers in the future. That same night,
Corliss pretends to be older than she is and starts flirting shamelessly with young
Private Jimmy Earhart, who has been invited into the Archer home for dinner. Corliss is
actually dating the boy next door, Dexter Franklin. Late that night, Lenny Archer returns
home on a short leave from his Air Force service. He goes directly to his girlfriend
Mildred and proposes to
</information>

think
To determine which government position was held by the woman who portrayed Corliss Archer
in the film Kiss and Tell, I will follow these steps:

1. Identify the woman who portrayed Corliss Archer in the film Kiss and Tell.
2. Find out which government position she held.

Let's start with step 1: identifying the woman who portrayed Corliss Archer in the film
Kiss and Tell.

<search>Corliss Archer Kiss and Tell </search><information>
Doc 1(Title: "Kiss and Tell
(play)") 1945. The play was produced by George Abbott and written by F. Hugh Herbert. It
ran for a total of 956 performances before closing on June 23, 1945. Two teenage girls
become interested in boys, and the girls' parents are making more problems rather than
solving them. A film version written by Herbert and based on his play was released by
Columbia Pictures on October 4, 1945 with Shirley Temple in the role of Corliss Archer.
The film was directed by Richard Wallace and produced by Sol C. Siegel. Temple would
reprise the role in the film's sequel ""A Kiss

Doc 2(Title: "Kiss and Tell (1945 film)") Kiss and Tell (1945 film) Kiss and Tell is a
1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. In
the film, two teenage girls cause their respective parents much concern when they start to
become interested in boys. The parents' bickering about which girl is the worse influence
causes more problems than it solves. The movie was based on the Broadway play ""Kiss and
Tell"", which was based on the Corliss Archer short stories. The stories, play and movie
were all written by F. Hugh Herbert. A sequel film, ""A Kiss for Corliss"", was released
in

Doc 3(Title: "Meet Corliss Archer") The program was produced by Ziv Productions. Several
episodes of the Ziv version are available on DVD, and some are also available from the
Internet Archive. Robin Morgan portrayed Corliss in a live telecast of ""Kiss and Tell""
on ""The Alcoa Hour"" (August 5, 1956), with Warren Berlinger as Dexter. Radio listeners
had to use their imaginations to visualize Corliss, her friends and her town. But those
imaginations got a boost in 1948 when the ""Meet Corliss Archer"" comic book, published
by Fox Feature Syndicate, came out in three issues from March to July 1948. Al Feldstein
(Albert B. Feldstein),

Doc 4(Title: "A Kiss for Corliss") A Kiss for Corliss A Kiss for Corliss is a 1949
American comedy film directed by Richard Wallace and written by Howard Dimsdale. It stars
Shirley Temple in her final starring role as well as her final film appearance. It is a
sequel to the 1945 film ""Kiss and Tell"". ""A Kiss for Corliss"" was retitled ""Almost a
Bride"" before release and this title appears in the title sequence. The film was released
on November 25, 1949, by United Artists. Corliss Archer (Shirley Temple) is an impulsive
teenager excited by notorious playboy Kenneth Marquis (David Niven), who has already been

Doc 5(Title: "Kiss and Tell (play)") Kiss and Tell (play) Kiss and Tell is a 1943
Broadway play by F. Hugh Herbert. ""Kiss and Tell"" starred Joan Caulfield as Corliss
Archer and Judith Parrish as her friend Mildred Pringle. The play's great success led to
offers from Hollywood for Caulfield, who left the production almost one year to the day
from when it opened, and was replaced in her role by her sister Betty Caulfield. ""Kiss
and Tell"" opened at the Biltmore Theatre in New York City on March 17, 1943. It remained
there until the end of 1944, before transferring to the Bijou Theatre in

Doc 6(Title: "Kiss and Tell (1945 film)") reported that the film hadn't changed much from
the ""protracted but reasonably diverting"" stage version, and singled out Jerome
Courtland as ""one of the most capable adolescent actors to come along in years."" Kiss
and Tell (1945 film) Kiss and Tell is a 1945 American comedy film starring then
17-year-old Shirley Temple as Corliss Archer. In the film, two teenage girls cause their
respective parents much concern when they start to become interested in boys. The parents'
bickering about which girl is the worse influence causes more problems than it solves. The
movie was based on the Broadway play ""Kiss

Doc 7(Title: "Kiss and Tell (play)") for Corliss"" in 1949. Kiss and Tell (play) Kiss and
Tell is a 1943 Broadway play by F. Hugh Herbert. ""Kiss and Tell"" starred Joan Caulfield
as Corliss Archer and Judith Parrish as her friend Mildred Pringle. The play's great
success led to offers from Hollywood for Caulfield, who left the production almost one
year to the day from when it opened, and was replaced in her role by her sister Betty
Caulfield. ""Kiss and Tell"" opened at the Biltmore Theatre in New York City on March 17,
1943. It remained there until the end of 1944, before transferring to
```

```
Doc 8(Title: "Meet Corliss Archer") of a series of stories. ""Kiss and Tell"" was a 1943
play that was adapted for a 1945 film starring Shirley Temple. The 1949 sequel, ""A Kiss
for Corliss"", was re-released in 1954. Like many other radio shows, ""Meet Corliss
Archer"" made the leap to television with live performances in 1951 and 1952, and from
1954 to 1955, as a syndicated television show starring Ann Baker and Mary Brian. One of
the show’s unique features was the occasional cut to a comic-book-style drawing, with
announcer’s commentary, that illustrated the current story situation and was used several
times during each episode.

Doc 9(Title: "Kiss and Tell (1945 film)") 1949 and also starred Temple, but was not
written by Herbert. To boost sales and attract customers at the local bazaar,
fifteen-year-old Corliss Archer and seventeen-year-old Mildred Pringle decide to start
selling kisses. When their booth at a USO bazaar fails to attract customers, teenager
Corliss Archer suggests to her best friend, Mildred Pringle, that they sell kisses. The
idea becomes a success among the soldiers visiting the bazaar, and business is booming,
until the girls’ mothers find out about it. Despite the fact that it was Corliss’ idea,
Mrs. Archer blames Mildred for the girls’ behavior. Mrs. Archer’s assumptions

Doc 10(Title: "Kiss and Tell (1945 film)") greatly upset Mr. Pringle when he hears about
them at the Pringle dinner table. He decides the family, including Mildred and their son
Raymond, will refrain from socializing with the Archers in the future. That same night,
Corliss pretends to be older than she is and starts flirting shamelessly with young
Private Jimmy Earhart, who has been invited into the Archer home for dinner. Corliss is
actually dating the boy next door, Dexter Franklin. Late that night, Lenny Archer returns
home on a short leave from his Air Force service. He goes directly to his girlfriend
Mildred and proposes to
</information>

(Title: "Kiss and Tell (1945 film)") Kiss and Tell (1945 film) Kiss and Tell is a 1945
American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. In the
film, two teenage girls cause their respective parents much concern when they start to
become interested in boys. The parents’ bickering about which girl is the worse influence
causes more problems than it solves. The movie was based on the Broadway play ""Kiss and
Tell"", which was based on the Corliss Archer short stories. The stories, play and movie
were all written by F. Hugh Herbert. A sequel film, ""A Kiss for Corliss"", was released
in</information>
(answer) Secretary of State

Therefore, the woman who portrayed Corliss Archer in the film Kiss and Tell held the
position of Secretary of State.

think
To determine which government position was held by the woman who portrayed Corliss Archer
in the film Kiss and Tell, I will follow these steps:

1. Identify the woman who portrayed Corliss Archer in the film Kiss and Tell.
2. Find out which government position she held.

Let’s start with step 1: identifying the woman who portrayed Corliss Archer in the film
Kiss and Tell.

<search>Corliss Archer Kiss and Tell</search><information>
Doc 1(Title: "Kiss and Tell
(play)") 1945. The play was produced by George Abbott and written by F. Hugh Herbert. It
ran for a total of 956 performances before closing on June 23, 1945. Two teenage girls
become interested in boys, and the girls’ parents are making more problems rather than
solving them. A film version written by Herbert and based on his play was released by
Columbia Pictures on October 4, 1945 with Shirley Temple in the role of Corliss Archer.
The film was directed by Richard Wallace and produced by Sol C. Siegel. Temple would
reprise the role in the film’s sequel ""A Kiss

Doc 2(Title: "Kiss and Tell (1945 film)") Kiss and Tell (1945 film) Kiss and Tell is a
1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. In
the film, two teenage girls cause their respective parents much concern when they start to
become interested in boys. The parents’ bickering about which girl is the worse influence
causes more problems than it solves. The movie was based on the Broadway play ""Kiss and
Tell"", which was based on the Corliss Archer short stories. The stories, play and movie
were all written by F. Hugh Herbert. A sequel film, ""A Kiss for Corliss"", was released
in

Doc 3(Title: "Meet Corliss Archer") The program was produced by Ziv Productions. Several
episodes of the Ziv version are available on DVD, and some are also available from the
Internet Archive. Robin Morgan portrayed Corliss in a live telecast of ""Kiss and Tell""
on ""The Alcoa Hour"" (August 5, 1956), with Warren Berlinger as Dexter. Radio listeners
had to use their imaginations to visualize Corliss, her friends and her town. But those
imaginations got a boost in 1948 when the ""Meet Corliss Archer"" comic book, published
by Fox Feature Syndicate, came out in three issues from March to July 1948. Al Feldstein
(Albert B. Feldstein),
```

```
Doc 4(Title: "A Kiss for Corliss") A Kiss for Corliss A Kiss for Corliss is a 1949
American comedy film directed by Richard Wallace and written by Howard Dimsdale. It stars
Shirley Temple in her final starring role as well as her final film appearance. It is a
sequel to the 1945 film ""Kiss and Tell"". ""A Kiss for Corliss"" was retitled ""Almost a
Bride"" before release and this title appears in the title sequence. The film was released
on November 25, 1949, by United Artists. Corliss Archer (Shirley Temple) is an impulsive
teenager excited by notorious playboy Kenneth Marquis (David Niven), who has already been

Doc 5(Title: "Kiss and Tell (play)") Kiss and Tell (play) Kiss and Tell is a 1943
Broadway play by F. Hugh Herbert. ""Kiss and Tell"" starred Joan Caulfield as Corliss
Archer and Judith Parrish as her friend Mildred Pringle. The play's great success led to
offers from Hollywood for Caulfield, who left the production almost one year to the day
from when it opened, and was replaced in her role by her sister Betty Caulfield. ""Kiss
and Tell"" opened at the Biltmore Theatre in New York City on March 17, 1943. It remained
there until the end of 1944, before transferring to the Bijou Theatre in

Doc 6(Title: "Kiss and Tell (1945 film)") reported that the film hadn't changed much from
the ""protracted but reasonably diverting"" stage version, and singled out Jerome
Courtland as ""one of the most capable adolescent actors to come along in years."" Kiss
and Tell (1945 film) Kiss and Tell is a 1945 American comedy film starring then
17-year-old Shirley Temple as Corliss Archer. In the film, two teenage girls cause their
respective parents much concern when they start to become interested in boys. The parents'
bickering about which girl is the worse influence causes more problems than it solves. The
movie was based on the Broadway play ""Kiss

Doc 7(Title: "Kiss and Tell (play)") for Corliss"" in 1949. Kiss and Tell (play) Kiss and
Tell is a 1943 Broadway play by F. Hugh Herbert. ""Kiss and Tell"" starred Joan Caulfield
as Corliss Archer and Judith Parrish as her friend Mildred Pringle. The play's great
success led to offers from Hollywood for Caulfield, who left the production almost one
year to the day from when it opened, and was replaced in her role by her sister Betty
Caulfield. ""Kiss and Tell"" opened at the Biltmore Theatre in New York City on March 17,
1943. It remained there until the end of 1944, before transferring to

Doc 8(Title: "Meet Corliss Archer") of a series of stories. ""Kiss and Tell"" was a 1943
play that was adapted for a 1945 film starring Shirley Temple. The 1949 sequel, ""A Kiss
for Corliss"", was re-released in 1954. Like many other radio shows, ""Meet Corliss
Archer"" made the leap to television with live performances in 1951 and 1952, and from
1954 to 1955, as a syndicated television show starring Ann Baker and Mary Brian. One of
the show's unique features was the occasional cut to a comic-book-style drawing, with
announcer's commentary, that illustrated the current story situation and was used several
times during each episode.

Doc 9(Title: "Kiss and Tell (1945 film)") 1949 and also starred Temple, but was not
written by Herbert. To boost sales and attract customers at the local bazaar,
fifteen-year-old Corliss Archer and seventeen-year-old Mildred Pringle decide to start
selling kisses. When their booth at a USO bazaar fails to attract customers, teenager
Corliss Archer suggests to her best friend, Mildred Pringle, that they sell kisses. The
idea becomes a success among the soldiers visiting the bazaar, and business is booming,
until the girls' mothers find out about it. Despite the fact that it was Corliss' idea,
Mrs. Archer blames Mildred for the girls' behavior. Mrs. Archer's assumptions

Doc 10(Title: "Kiss and Tell (1945 film)") greatly upset Mr. Pringle when he hears about
them at the Pringle dinner table. He decides the family, including Mildred and their son
Raymond, will refrain from socializing with the Archers in the future. That same night,
Corliss pretends to be older than she is and starts flirting shamelessly with young
Private Jimmy Earhart, who has been invited into the Archer home for dinner. Corliss is
actually dating the boy next door, Dexter Franklin. Late that night, Lenny Archer returns
home on a short leave from his Air Force service. He goes directly to his girlfriend
Mildred and proposes to
</information>

(Title: "Kiss and Tell (1945 film)") Kiss and Tell (1945 film) Kiss and Tell is a 1945
American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. In the
film, two teenage girls cause their respective parents much concern when they start to
become interested in boys. The parents' bickering about which girl is the worse influence
causes more problems than it solves. The movie was based on the Broadway play ""Kiss and
Tell"", which was based on the Corliss Archer short stories. The stories, play and movie
were all written by F. Hugh Herbert. A sequel film, ""A Kiss for Corliss"", was released
in </information>
(answer) Secretary of State

Therefore, the woman who portrayed Corliss Archer in the film Kiss and Tell held the
position of Secretary of State.

think
To determine which government position was held by the woman who portrayed Corliss Archer
in the film Kiss and Tell, I will follow these steps:

1. Identify the woman who portrayed Corliss Archer in the film Kiss and Tell.
```

```
2. Find out which government position she held.

Let’s start with step 1: identifying the woman who portrayed Corliss Archer in the film
Kiss and Tell.

<search>Corliss Archer Kiss and Tell</search><information>
Doc 1(Title: "Kiss and Tell
(play)") 1945. The play was produced by George Abbott and written by F. Hugh Herbert. It
ran for a total of 956 performances before closing on June 23, 1945. Two teenage girls
become interested in boys, and the girls’ parents are making more problems rather than
solving them. A film version written by Herbert and based on his play was released by
Columbia Pictures on October 4, 1945 with Shirley Temple in the role of Corliss Archer.
The film was directed by Richard Wallace and produced by Sol C. Siegel. Temple would
reprise the role in the film’s sequel ""A Kiss

Doc 2(Title: "Kiss and Tell (1945 film)") Kiss and Tell (1945 film) Kiss and Tell is a
1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. In
the film, two teenage girls cause their respective parents much concern when they start to
become interested in boys. The parents’ bickering about which girl is the worse influence
causes more problems than it solves. The movie was based on the Broadway play ""Kiss and
Tell"", which was based on the Corliss Archer short stories. The stories, play and movie
were all written by F. Hugh Herbert. A sequel film, ""A Kiss for Corliss"", was released
in

Doc 3(Title: "Meet Corliss Archer") The program was produced by Ziv Productions. Several
episodes of the Ziv version are available on DVD, and some are also available from the
Internet Archive. Robin Morgan portrayed Corliss in a live telecast of ""Kiss and Tell""
on ""The Alcoa Hour"" (August 5, 1956), with Warren Berlinger as Dexter. Radio listeners
had to use their imaginations to visualize Corliss, her friends and her town. But those
imaginations got a boost in 1948 when the ""Meet Corliss Archer"" comic book, published
by Fox Feature Syndicate, came out in three issues from March to July 1948. Al Feldstein
(Albert B. Feldstein),

Doc 4(Title: "A Kiss for Corliss") A Kiss for Corliss A Kiss for Corliss is a 1949
American comedy film directed by Richard Wallace and written by Howard Dimsdale. It stars
Shirley Temple in her final starring role as well as her final film appearance. It is a
sequel to the 1945 film ""Kiss and Tell"". ""A Kiss for Corliss"" was retitled ""Almost a
Bride"" before release and this title appears in the title sequence. The film was released
on November 25, 1949, by United Artists. Corliss Archer (Shirley Temple) is an impulsive
teenager excited by notorious playboy Kenneth Marquis (David Niven), who has already been

Doc 5(Title: "Kiss and Tell (play)") Kiss and Tell (play) Kiss and Tell is a 1943
Broadway play by F. Hugh Herbert. ""Kiss and Tell"" starred Joan Caulfield as Corliss
Archer and Judith Parrish as her friend Mildred Pringle. The play’s great success led to
offers from Hollywood for Caulfield, who left the production almost one year to the day
from when it opened, and was replaced in her role by her sister Betty Caulfield. ""Kiss
and Tell"" opened at the Biltmore Theatre in New York City on March 17, 1943. It remained
there until the end of 1944, before transferring to the Bijou Theatre in

Doc 6(Title: "Kiss and Tell (1945 film)") reported that the film hadn’t changed much from
the ""protracted but reasonably diverting"" stage version, and singled out Jerome
Courtland as ""one of the most capable adolescent actors to come along in years."" Kiss
and Tell (1945 film) Kiss and Tell is a 1945 American comedy film starring then
17-year-old Shirley Temple as Corliss Archer. In the film, two teenage girls cause their
respective parents much concern when they start to become interested in boys. The parents’
bickering about which girl is the worse influence causes more problems than it solves. The
movie was based on the Broadway play ""Kiss

Doc 7(Title: "Kiss and Tell (play)") for Corliss"" in 1949. Kiss and Tell (play) Kiss and
Tell is a 1943 Broadway play by F. Hugh Herbert. ""Kiss and Tell"" starred Joan Caulfield
as Corliss Archer and Judith Parrish as her friend Mildred Pringle. The play’s great
success led to offers from Hollywood for Caulfield, who left the production almost one
year to the day from when it opened, and was replaced in her role by her sister Betty
Caulfield. ""Kiss and Tell"" opened at the Biltmore Theatre in New York City on March 17,
1943. It remained there until the end of 1944, before transferring to

Doc 8(Title: "Meet Corliss Archer") of a series of stories. ""Kiss and Tell"" was a 1943
play that was adapted for a 1945 film starring Shirley Temple. The 1949 sequel, ""A Kiss
for Corliss"", was re-released in 1954. Like many other radio shows, ""Meet Corliss
Archer"" made the leap to television with live performances in 1951 and 1952, and from
1954 to 1955, as a syndicated television show starring Ann Baker and Mary Brian. One of
the show’s unique features was the occasional cut to a comic-book-style drawing, with
announcer’s commentary, that illustrated the current story situation and was used several
times during each episode.

Doc 9(Title: "Kiss and Tell (1945 film)") 1949 and also starred Temple, but was not
written by Herbert. To boost sales and attract customers at the local bazaar,
fifteen-year-old Corliss Archer and seventeen-year-old Mildred Pringle decide to start
selling kisses. When their booth at a USO bazaar fails to attract customers, teenager
```

```
Corliss Archer suggests to her best friend, Mildred Pringle, that they sell kisses. The
idea becomes a success among the soldiers visiting the bazaar, and business is booming,
until the girls' mothers find out about it. Despite the fact that it was Corliss' idea,
Mrs. Archer blames Mildred for the girls' behavior. Mrs. Archer's assumptions

Doc 10(Title: "Kiss and Tell (1945 film)") greatly upset Mr. Pringle when he hears about
them at the Pringle dinner table. He decides the family, including Mildred and their son
Raymond, will refrain from socializing with the Archers in the future. That same night,
Corliss pretends to be older than she is and starts flirting shamelessly with young
Private Jimmy Earhart, who has been invited into the Archer home for dinner. Corliss is
actually dating the boy next door, Dexter Franklin. Late that night, Lenny Archer returns
home on a short leave from his Air Force service. He goes directly to his girlfriend
Mildred and proposes to
</information>

(  the My the   Hugh of kiss on but years British</ Mr years</ years think</K Archer
Archer</</ Shirley version</ the</ Q</My</</</</</</</</  woman</ years film, with ""</
years</ ""</ the Secretary</ Shirley</</ ""</</    a L Shirley</ Cor "" Jerome>
years</Cor</ Cor</ Cor</ Cor previous</ My with from my a a more, To  woman</  [ more</</
action my the top</ film</ Secretary</ My information</night and Kiss</ The the
Secretary with the Kiss is lion Q the previous: "" Shirley with boy ""</</ with the top
the Scott </ the Cor</""</. "" Archer</ Cor</ Cor</ Secretary</ and the a the</ Cor, Cor
with the Secretary</</</ Kiss</</ Shirley</previous</ My. Cor seventeen. The girl Cor,
The7</  1</</ My answer with the ""</ L Myiss</ The with Dexter, From. My</ The</ Private
</ Cor about</ Cor</ My system</</ Cor. a The</ Jerome. My a</My Secretary Cor</ Cor: The
film and the Cor</ My compile
 < </</ Kiss with the think kiss</ Kiss</</</ information, The  </ Previous My</ the
My</ Secretarylaction forCor</ Secretaryiss Q</ Scott</ My</   system in the My  The
Secretary</ the teenage, womans| My</</</</ < a</
 < years</</ The tell, top Dexter, The film with the Tell</</ system. Hugh Cor</ ""</
The the the</ < the Broadway Q Cor government Q the boys and X Cor Archer</ [</ is action
Kiss</ the answer Richard Kiss></  top Kiss></ the So
 My</</ with the the the title "" the the them the previous Kiss </ kiss</ My previous
previous</ "" the</ Cor action</ Cor</</ the top Cor Cor</ < the ""</ ""</</my Cor
which</ The( < The the girls>not the the a</ ""</ Cor government Cor</ Cor)the</ Cor boys
Q</  Secretary</</""</ the her years the </ Archer</  </ my the stories. My their</
">Cor</ a the the girls</</ Joan "" years</</ Secretary of the the a parents</</ Cor
Cor the the the a Kiss</</ information</""pro with the</doc</ Coriss</ the me previous
were new</15 the{>""> if Secretary Tell</</""Cor</ more</ Cor</</</</</</ ""</</>

Corlovernmentliss</K the stage Corissed</ is is F </ the</ " "" the and is</ that</ Tell
and My Shirley: My the my</ the the years is a The (My Archer</information, IMy</{ My of
a""</ Kiss</ Kiss think years and Secretary of</ Myl the film Herbert, My government your
film "" the top theant. F</ Q</ "" Archer of my my think</ new""> which{K think</ <
action</</ Q the the answer to I. The</ Q action Q the</ my think Q Cor Cor the the the
film was, The, My Cor and I Kiss</ Cor my some the  the</ My</ My my with the the My the
film</ Secretary</ Secretary and theiriss</ Cor and best my the the The</ Cor the her
Dextery for "" Secretary by woman,Secretary Dexter book Mild as> the  first the them</
Secretary. film, My</ *</, My previous is a "" </ a the My</ the Al the Sol</</  Cor</ a I
My</ My my with the the My the the</ with with stage the a the Cor: Cor: My ""ensored</
Secretary O years Secretary"", "" the the the the the the the Shirley></</ is the</ title,
The government</ and with with Shirley Archer. My</</ Cor"", and</ Cor</</ Cor not My </
</></</ . the</ action [ with think by answer Cor think "" in government</ "" with the
years with the previous action"".""</ < ( most "" the query, Which the answer is provide</
invalid "" with woman"" Kiss</ mild the a girl</ play</ Kiss, ( movie "" the the <
fromifld answer</ the more State between</</ Secretary. < Archer</ with kiss</ Kiss
normal Mr</ My Cor</ Kiss</.</</, My "", Based Cor</ Cor "" the</ L</</</ Cor previous
previous to My my</ Myiss Mc position in My</ information with</</</</</ Kississar the</
Cor</ Corfalyn Cor</ My, My in, My the that the previous the a the a the previous Dexter
Court in a a a not</with that film Cor</ actionliss Archer at the stage ""</ Cor</</
action</ </, which 1 government</ a theThe government in a film</ Cor Archer</ the movie</
information</ as nighticatein query with was the a when ""</ rooms andrown</ years ""</</
night</"" Cor</"</" action</,</</ with them the a"" night</actioniss</ Cor</"" Corliss
Archer Position the Cor Kiss with the it the "" tell ""s</ years Kiss that all night</</
Kiss</ Cory, and that that as with the the. with</ Kiss their</"" Archer "" Kiss """"</
Kiss.K night years</ action</ Kiss</</ " and9 The Kiss</ Kiss and b−nightiss</</iss a</ a
</""amy</ search</ "" K previous corp</s the a Kiss My the the the same night kiss. ""
afilm the a oneparent theparent previous; L</</</film Archer, a the the film,   andiss</
this the them with My previous the My ArcherCor, The</ "" that the</My a a−year the
previous My the Kiss</My</  My a My Cor ""pro two the Archer</L Cor/""</"> the Archer</
Kiss withteen the Scott the the teenager the the woman, film. The years"" the the the the
a 7</ the My previous in Cor the My two the My the boys Cor Cor the the the with two, the
the the The film</ the the the The</ Cor
Cor the of the short</ Cor the</Cor the</ the the</ Cor the and the the</ the</> the the
the</</ the</ "   </ Tell</ Cor best. A < in localss</ ""</""</ the</ Kiss</ Cor</ Cor
the the andcor my</ Dexter, as Kiss, Dexter</</My Cor Kiss</ Archer"" Dexter the</ ""</>
Cor</</ Cor</</ Cor</</ My years the a a the I the more Cor</ Archer</ kiss</ Scott"""",
with young, Cor</ Kiss</</"" them"","", Private"" Jimmy, Shirley""rok the a  Cor) Scott
with young</</ Cor""K"" * a Kiss)</ with night their</  Cor</ Cor</>
```

```
Dexter:before</ June, The invalid invalid</ kiss""</.
 My</ The the</</</</ more years.</</</</ Cor Cor. valid</ Cor Cor the kisstd new aly a
the woman, My My me</ </ kiss</ my the topic Cor who</ most Cor Cor </</ </ information</
My</ illustration</ My new</</ action</ </ My previous</My, could the <  Kiss</ the think
 </.,ant."" kississ</ query</ xx tell. same </ tell</) in kiss</ kiss</ Jerome kiss with
the a some</ assumptions</ and. query</K1, that Title:</ previous is as ( invalid</
invalid all. My My the the or note off the 1</ Kiss to </  Cor</ Kiss</"" their with the
a the a</</ with the   woman, My previous Cor</ Cor much "" him "" Kiss"" " the the but
Dexter Tell</ This  </ a  action</ Kiss Tell</</</</ I</</ is a the</ Cor the theMy</ his
his their</ westerns> Cor as Cor</</</</</ Dexter</information Cor</think "" as</ Kiss</
the Scott</ one</</</ my</</ an the </  is the Archer</</  a a""</ which two is the
query</ </ </ Archer</ June</for) those</ more my,  my a My night</ Shirley</ Kiss</ the
answer Cor</ My Secretary</ Cor action Jerome</Mr with the Scott</ Dexter</</ Cor</</ Cor
</</</</</ with Dexter Cor the and</</ Cor</</ Cor></"""></ Cor"""" Hugh</ with</</</</
</ Secretary "" the the</Secretary</ Cor</ government</ Secretary</</</</s Cor government
</ with</</ with</</ with</</ the</</ Cor</,</ with</</</s</</</</</ Kiss</</</</</</</</
</ Kiss</
</</</</</</</</ Cor</ Cor Cor</ Cor</ Cor</</cor

 and kiss</</ Cor</ Cor the a a</, Dexter</""</ Dexter</</</
```

Prompt: Search-Augmented Reasoning

```
[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}
```

Prompt: Control Messages

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

<control>Do not expand any retrieved documents. Continue the search.</control>

<control>Stop searching</control>

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

Example: 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?


[ASSISTANT 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>


[ASSISTANT 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>


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

Example: 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?


[ASSISTANT 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>


[ASSISTANT 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>
```

```
[ASSISTANT 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>


[ASSISTANT 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>


[ASSISTANT 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>


[ASSISTANT 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>


[ASSISTANT 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>


[ASSISTANT 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>
```

Example: 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?


[ASSISTANT 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>


[ASSISTANT 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>


[ASSISTANT 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>


[USER PROMPT]
<control>Expand the retrieved documents: [1911155, 13506636]</control>


[ASSISTANT 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"
NNigel 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>


[ASSISTANT 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>


[ASSISTANT 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>
```

Example: Search-Augmented Reasoning – HotpotQA (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: What county in Kentucky has a population over 43,000 and has Kentucky Route 46
run through it?


[ASSISTANT RESPONSE]
<think>
To find the county in Kentucky that has a population over 43,000 and has Kentucky Route
46 run through it, I should first search for Kentucky counties and their populations.
</think>
<search>Kentucky counties with population over 43,000 and Kentucky Route 46</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=17486196 score=0.8561
Title: Kentucky Route 47
Snippet: through more rural areas before heading to the north into forests and intersect-
ing the eastern terminus of KY 1112. KY 47 passes through more farm fields and woodland
before crossing a CSX railroad line and heading into ...

Rank=2 doc_id=17486194 score=0.8549
Title: Kentucky Route 47
Snippet: Kentucky Route 47 Kentucky Route 47 is a 9.851-mile state highway in Kentucky
that runs from KY 36 and KY 467 in Sanders to US 42 in Ghent KY 47 mostly stays in Carroll
County but makes a short entrance into Gallatin Cou...

Rank=3 doc_id=3091540 score=0.8527
Title: Kentucky Route 43
Snippet: Kentucky Route 43 Kentucky Route 43 is a state highway in Kentucky that runs from
BUS KY 55 and Boone Station Road northeast of Shelbyville to US 421 southeast of
Pleasureville. ...

Rank=4 doc_id=3091542 score=0.8524
Title: Kentucky Route 43
Snippet: through more rural areas before it ends at an intersection with US 421. Kentucky
Route 43 Kentucky Route 43 is a state highway in Kentucky that runs from BUS KY 55 and
Boone Station Road northeast of Shelbyville to US 421 southeast of Pleasureville. ...

Rank=5 doc_id=319230 score=0.8509
Title: U.S. Route 45 in Kentucky
Snippet: turns to the northwest and ends at a junction with US 45 south of the Brookport
Bridge over the Ohio River from its origin. ...
</search_results>


[USER PROMPT]
<control>Do not expand any retrieved documents. Continue the search.</control>


[ASSISTANT RESPONSE]
<think>
I have found some results related to Kentucky counties with population over 43,000 and the
routes Kentucky Route 47 and Kentucky Route 43. However, none of the snippets clearly
mention a county with a population over 43,000 that has Kentucky Route 46 running through
it. I need to search for information specifically related to Kentucky Route 46 and
counties.
</think>
<search>Kentucky Route 46 and counties with population over 43,000</search>


[USER PROMPT]
<search_results>
```

```
Rank=1 doc_id=17486196 score=0.8631
Title: Kentucky Route 47
Snippet: through more rural areas before heading to the north into forests and intersect-
ing the eastern terminus of KY 1112. KY 47 passes through more farm fields and woodland
before crossing a CSX railroad line and heading into ...

Rank=2 doc_id=17486194 score=0.8613
Title: Kentucky Route 47
Snippet: Kentucky Route 47 Kentucky Route 47 is a 9.851-mile state highway in Kentucky
that runs from KY 36 and KY 467 in Sanders to US 42 in Ghent KY 47 mostly stays in Carroll
County but makes a short entrance into Gallatin Cou...

Rank=3 doc_id=319230 score=0.8572
Title: U.S. Route 45 in Kentucky
Snippet: turns to the northwest and ends at a junction with US 45 south of the Brookport
Bridge over the Ohio River from its origin. ...

Rank=4 doc_id=7320080 score=0.8562
Title: Kentucky Route 46
Snippet: Kentucky Route 46 Kentucky Route 46 (KY 46) is a state highway in Nelson County,
Kentucky. KY 46 runs from KY 52 south of Nelsonville to KY 49 south of Greenbrier. ...

Rank=5 doc_id=3091540 score=0.8549
Title: Kentucky Route 43
Snippet: Kentucky Route 43 Kentucky Route 43 is a state highway in Kentucky that runs from
BUS KY 55 and Boone Station Road northeast of Shelbyville to US 421 southeast of
Pleasureville. ...
</search_results>


[USER PROMPT]
<control>Expand the retrieved documents: [7320080]</control>


[ASSISTANT RESPONSE]
<think>
The search results found show that Kentucky Route 46 runs through Nelson County, and the
snippet mentions that it ends at KY 49 south of Greenbrier. However, the snippet does not
explicitly state the population of Nelson County or whether it has a population over
43,000. I should expand the information about Nelson County to find the required details.
</think>
<expand>{"doc_ids": ["7320080"]} </expand>


[USER PROMPT]
<information>
[DOC 7320080]
"Kentucky Route 46"
Kentucky Route 46 Kentucky Route 46 (KY 46) is a state highway in Nelson County,
Kentucky. KY 46 runs from KY 52 south of Nelsonville to KY 49 south of Greenbrier. KY 46
heads east from KY 52 as Nat Rogers Road. About northwest of Culvertown, KY 46 begins
heading in a southeast direction. At Culvertown, KY 46 meets U.S. Route 31E (US 31E) and
joins the route heading northbound. KY 46 then branches off US 31E at Balltown. KY 46
heads due south for , before heading southeast. KY 46 continues for about before
terminating at KY 49 about
</information>


[ASSISTANT RESPONSE]
<think>
The information found mentions that Kentucky Route 46 runs through Nelson County,
Kentucky. However, there is no explicit mention of the population of Nelson County. To
proceed, I will need to search for the population of Nelson County, Kentucky.
</think>
<search>Nelson County, Kentucky population</search>


[USER PROMPT]
<search_results>
Rank=1 doc_id=21001401 score=0.8752
Title: Nelson County, Kentucky
Snippet: created Virginia county was named in his honor. According to the U.S. Census
Bureau, the county has a total area of , of which is land and (1. ...

Rank=2 doc_id=21001402 score=0.8654
Title: Nelson County, Kentucky
Snippet: the population were Hispanics or Latinos of any race. There were 16,826
households out of which 36.60% had children under the age of 18 living with them, 52. ...
```

```
Rank=3 doc_id=21001400 score=0.8611
Title: Nelson County, Kentucky
Snippet: Nelson County, Kentucky Nelson County is a county located in the U.S. state of
Kentucky. As of the 2010 census, the population was 43,437. ...

Rank=4 doc_id=21001403 score=0.8550
Title: Nelson County, Kentucky
Snippet: 18 to 24, 26.47% from 25 to 44, 27.84% from 45 to 64, and 11.73% who were 65 or
older. ...

Rank=5 doc_id=21001404 score=0.8429
Title: Nelson County, Kentucky
Snippet: for the county was $18,120. About 10.00% of families and 12.20% of the population
were below the poverty line, including 15.
</search_results>


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


[ASSISTANT RESPONSE]
<think>
The control message says 'Stop searching', so I cannot use search or expand. I must
answer based on the available information.

The search results provide population data for Nelson County, Kentucky. According to the
U.S. Census Bureau, as of the 2010 census, Nelson County had a population of 43,437. This
population exceeds 43,000, indicating that Nelson County meets the criteria of having a
population over 43,000.

The snippet also does not specify whether Kentucky Route 46 runs through Nelson County.
However, based on previous search results, we know that Kentucky Route 46 runs through
Nelson County, Kentucky.

Therefore, I can conclude that the county in Kentucky with a population over 43,000 and
which Kentucky Route 46 runs through is Nelson County.
</think>
<answer>Nelson County</answer>
```