

Chaining the Evidence: Robust Reinforcement Learning for Deep Search Agents with Citation-Aware Rubric Rewards

Jiajie Zhang^{1*}, Xin Lv², Ling Feng¹, Lei Hou¹, Juanzi Li¹

¹Tsinghua University, ²Zhipu AI

Abstract

Reinforcement learning (RL) has emerged as a critical technique for enhancing LLM-based deep search agents. However, existing approaches primarily rely on binary outcome rewards, which fail to capture the comprehensiveness and factuality of agents' reasoning process, and often lead to undesirable behaviors such as shortcut exploitation and hallucinations. To address these limitations, we propose **Citation-aware Rubric Rewards (CaRR)**, a fine-grained reward framework for deep search agents that emphasizes reasoning comprehensiveness, factual grounding, and evidence connectivity. CaRR decomposes complex questions into verifiable single-hop rubrics and requires agents to satisfy these rubrics by explicitly identifying hidden entities, supporting them with correct citations, and constructing complete evidence chains that link to the predicted answer. We further introduce **Citation-aware Group Relative Policy Optimization (C-GRPO)**, which combines CaRR and outcome rewards for training robust deep search agents. Experiments show that C-GRPO consistently outperforms standard outcome-based RL baselines across multiple deep search benchmarks. Our analysis also validates that C-GRPO effectively discourages shortcut exploitation, promotes comprehensive, evidence-grounded reasoning, and exhibits strong generalization to open-ended deep research tasks. Our code and data are available at <https://github.com/THUDM/CaRR>.

1 Introduction

Recently, LLM-based deep search agents have attracted growing attention for their ability to leverage external web-browsing tools to solve complex, knowledge-intensive problems (Yao et al., 2023; Wang et al., 2024; OpenAI, 2025a). A prominent

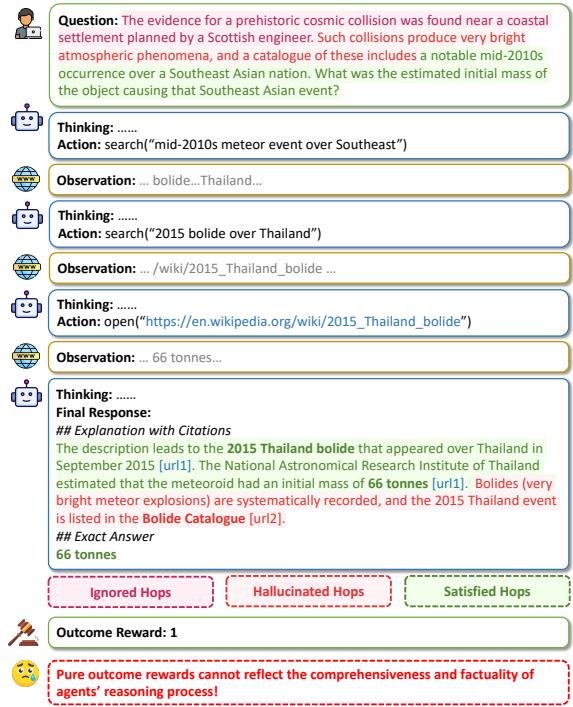


Figure 1: Pure outcome rewards fail to capture shortcut exploitation and hallucinations of deep search agents.

line of research has focused on applying reinforcement learning (RL) to further enhance these agents' long-horizon information-seeking capacity in the vast and noisy web environment, typically leveraging synthetic multi-hop QA datasets that are intentionally challenging but feature short-form answers for easy verification (Gao et al., 2025; Wu et al., 2025a; Li et al., 2025b; Lu et al., 2025). For the efficiency and scalability of RL, existing works commonly use only outcome rewards in training, which are binary signals indicating whether the agent's predicted final answer matches the ground truth (Jin et al., 2025a; Gao et al., 2025; Li et al., 2025b; Liu et al., 2025b).

While these outcome-based RL methods have demonstrated notable gains (Li et al., 2025a; Team et al., 2025), they suffer from inherent limitations.

*Work was done when JZ interned at Zhipu AI.

As illustrated in Figure 1, binary outcome rewards alone cannot accurately reflect the comprehensiveness and factuality of agents’ reasoning processes (Shao et al., 2025b), leaving room for undesirable behaviours: Agents may arrive at the correct answer by shortcut solutions (e.g., exploiting only a few hops of information while ignoring other constraints in the question) or fortunate hallucination. Optimizing toward these flawed trajectories will result in deep search agents with diminished robustness and suboptimal performance.

To address these limitations, we propose **Citation-aware Rubric Rewards (CaRR)**, a novel fine-grained reward framework for deep search agents that emphasizes reasoning comprehensiveness, factual grounding, and evidence connectivity. Our framework is inspired by the observation that each hop within the synthetic complex question can naturally serve as a checkpoint for evaluating the agent’s reasoning process: An ideal trajectory that completely solves the given question should satisfy all hops by revealing the identities of all intermediate hidden entities and supporting them with correct citations. Building upon this idea, our framework first employs an LLM to decompose the multi-hop question into a list of single-hop factual statements, each involves several hidden entities that should be found during exploration. These factual statements are then used as point-wise rubrics to assess the comprehensiveness and factuality of agents’ trajectories. Specifically, a rubric is satisfied by a trajectory only if (1) the identities of all relevant hidden entities are explicitly revealed in the final response; (2) the factual statement, along with the identified entities, is fully supported by the cited web contents; (3) the supported rubric can be connected to the predicted final answer via other supported rubrics, thereby constituting a complete evidence chain. Given a trajectory, we employ a judge LLM to check whether each rubric is satisfied following the above three criteria, and the citation-aware rubric reward is defined as the ratio of satisfied rubrics.

Building on CaRR, we further introduce **Citation-aware Group Relative Policy Optimization (C-GRPO)**, an extension of GRPO (Shao et al., 2024) that incorporates context-aware rubric rewards with traditional outcome rewards in RL. Specifically, C-GRPO assigns an additional weighted rubric reward to the trajectories whose outcome reward is 1. By doing so, C-GRPO preserves the primary objective of finding the correct

answer while encouraging the agent to produce more comprehensive and evidence-grounded reasoning processes, thereby achieving robust RL and better final performance.

To validate the efficacy of CaRR and C-GRPO, we conduct RL experiments on both small (4B) and large (30B) model scales. The evaluation results on four challenging deep search benchmarks indicate that C-GRPO consistently outperforms the GRPO baseline that uses pure outcome rewards, and also demonstrates significantly better performance when provided with extended context budgets. Our analysis reveals that C-GRPO successfully discourages shortcut exploitation and promotes more comprehensive, citation-supported solutions, yielding robust policies featured by rigorous self-verification and better factuality. Moreover, the agents trained with C-GRPO and synthetic QA data also generalize well on open-ended deep research tasks, even surpassing some advanced agents trained with proprietary data.

In summary, our main contributions include: (1) We identify key limitations of outcome-based RL in training deep search agents, including shortcut exploitation and hallucination tolerance; (2) We propose CaRR, a novel framework that provides fine-grained rewards for assessing the comprehensiveness and factuality of deep search agents; (3) We propose C-GRPO, a mixed-reward RL algorithm combining outcome rewards and context-aware rubric rewards for training robust deep search agents; (4) We conduct extensive experiments and thorough analysis to validate the efficacy of CaRR and C-GRPO.

2 Methodology

In this section, we first provide a brief overview of key concepts in deep search agents, then introduce our CaRR framework and C-GRPO algorithm.

2.1 Preliminary

Deep Search Agents. We adopt the ReAct (Yao et al., 2023) paradigm for deep search agents. Given a question, the LLM-based agent follows an iterative cycle of thinking, action (i.e., a tool call), and observation until obtaining the final answer. A complete trajectory with T iterations can be formalized as:

$$\mathcal{H} = (\tau_1, a_1, o_1, \dots, \tau_t, a_t, o_t, \dots, \tau_T, a_T), \quad (1)$$

where τ_t , a_t , o_t denote the thought, action, and observation at step t . Specifically, the action a_t

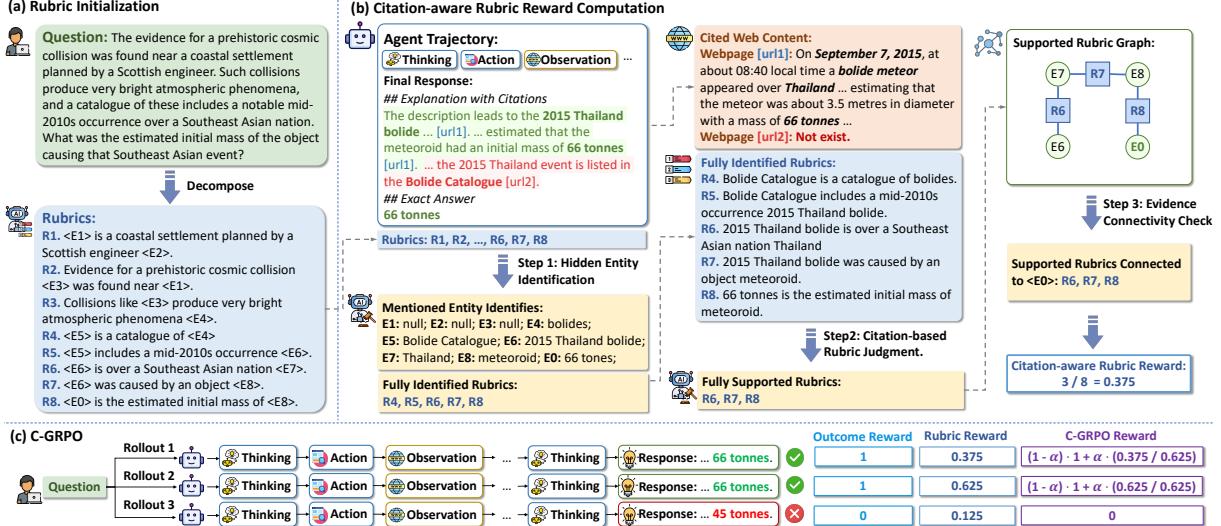


Figure 2: Overview of (a) rubric initialization; (b) computation of context-aware rubric rewards; (c) C-GRPO.

$(1 \leq t < T)$ calls one of the following three browsing tools: (1) a search tool that retrieves top- n relevant webpages for the given query and returns the title, URL, and snippet of each webpage; (2) an open tool that accesses the given URL and shows the head part of the page; (3) a find tool that matches the given keyword in the opened webpage and returns surrounding content of each match. While a_T is the final response, consisting of an explanation with citations and the final answer. The tool descriptions and trajectory format are detailed in Appendix A.

Synthetic Deep Search Training Data. RL of deep search agents typically relies on synthetic complex QA datasets (Gao et al., 2025; Li et al., 2025b; Lu et al., 2025). These datasets are commonly constructed from entity-centric knowledge graphs, involving multi-hop reasoning paths and deliberate information obfuscation to increase search complexity. For training convenience, the final answer is often a short-form entity string, allowing automatic correctness verification.

2.2 Citation-Aware Rubric Rewards

To address the limitations of outcome rewards, we propose **Citation-aware Rubric Rewards (CaRR)**, a novel fine-grained reward framework for deep search agents, taking into account reasoning comprehensiveness, factual grounding, and evidence connectivity. Specifically, CaRR utilizes the underlying compositional structure of the synthetic data. As illustrated in Figure 2, CaRR first decomposes a synthetic multi-hop question into a list of

atomic factual statements, each involves several hidden entities that need to be found. These atomic statements can naturally serve as point-wise, verifiable rubrics for assessing reasoning comprehensiveness and factuality of deep search agents: An ideal trajectory that completely solves the question should satisfy all rubrics by revealing the identities of corresponding hidden entities, supporting them with cited web contents, and connecting the supported rubrics to form complete evidence chains that link to the final answer. Moreover, the identified hidden entities and cited URLs should be detailed in the final response provided to the user. Based on this idea, CaRR uses a three-step method after the rubric initialization to provide fine-grained reward for agent rollouts, including: (1) hidden entity identification; (2) citation-based rubric judgment; and (3) evidence connectivity check. We will detail the rubric initialization and the three-step reward computation as follows.

2.2.1 Rubric Initialization

For each question q in the training set, we prompt an LLM $\mathcal{M}_{\text{rubric}}$ to decompose the question to locate hidden entities \mathcal{E}_q (i.e., entities that should be found when solving q) and generate the initial rubrics \mathcal{R}_q :

$$\mathcal{E}_q, \mathcal{R}_q = \mathcal{M}_{\text{rubric}}(q), \quad (2)$$

where

$$\begin{aligned} \mathcal{E}_q &= \{e_0, e_1, \dots, e_{n_q}\}, \\ \mathcal{R}_q &= \{r_1, \dots, r_{m_q}\}. \end{aligned} \quad (3)$$

As illustrated in Figure 2, each hidden entity $e_i \in \mathcal{E}_q$ is denoted by a placeholder <Ei>, and e_0 refers to the final answer. Each rubric $r_j = (s_j, \mathcal{E}_{q,j}) \in \mathcal{R}_q$ is an atomic factual statement s_j about an entity set $\mathcal{E}_{q,j} \subseteq \mathcal{E}_q$, and will serve as a checkpoint for assessing search agent's trajectories. Note that these rubrics are pre-generated before training and remain unchanged throughout the RL process.

2.2.2 Reward Computation

After initializing the hidden entity set \mathcal{E}_q and initial rubrics \mathcal{R}_q for a question q , given an agent trajectory $\mathcal{H} = (\tau_1, a_1, o_1, \dots, \tau_T, a_T)$, we use a three-step procedure with a judge LLM $\mathcal{M}_{\text{judge}}$ to assign a fine-grained rubric reward for \mathcal{H} , taking into account reasoning comprehensiveness, citation grounding, and evidence connectivity:

Step 1: Hidden Entity Identification. From the perspective of reasoning comprehensiveness, an ideal trajectory for solving q should consider all rubrics implied by q , uncover the identities of corresponding hidden entities during exploration, and explain them in the final response a_T . In light of this, we first employ $\mathcal{M}_{\text{judge}}$ to judge whether a_T explicitly identifies the name of each hidden entity $e_i \in \mathcal{E}_q$:

$$\{e_0^{\mathcal{H}}, \dots, e_n^{\mathcal{H}}\} = \mathcal{M}_{\text{judge}}(q, \mathcal{R}_q, \mathcal{E}_q, a_T), \quad (4)$$

where $e_i^{\mathcal{H}}$ is either the mentioned name of e_i in a_T ¹, or null if the name is not explicitly identified. Only rubrics whose hidden entities are all identified are regarded as being fully identified by \mathcal{H} . Formally, by defining the mapping:

$$f^{\mathcal{H}}(e_i) = e_i^{\mathcal{H}}, \quad \forall e_i \in \mathcal{E}_q \quad (5)$$

we instantiate each $r_j = (s_j, \mathcal{E}_{q,j}) \in \mathcal{R}_q$ by replacing hidden entities $\mathcal{E}_{q,j}$ with their identified name:

$$\begin{aligned} \mathcal{E}_{q,j}^{\mathcal{H}} &= \{f^{\mathcal{H}}(e_i) \mid e_i \in \mathcal{E}_{q,j}\}, \\ r_j^{\mathcal{H}} &= (s_j, \mathcal{E}_{q,j}^{\mathcal{H}}), \\ \mathcal{R}_q^{\mathcal{H}} &= \{r_1^{\mathcal{H}}, \dots, r_{m_q}^{\mathcal{H}}\}. \end{aligned} \quad (6)$$

Then the fully-identified rubrics are defined as:

$$\mathcal{R}_q^{\text{identify}} = \{r_j^{\mathcal{H}} \in \mathcal{R}_q^{\mathcal{H}} \mid e_i^{\mathcal{H}} \neq \text{null}, \forall e_i^{\mathcal{H}} \in \mathcal{E}_{q,j}^{\mathcal{H}}\}, \quad (7)$$

which will be selected for further judgment.

¹Note that we do not require the identified $e_i^{\mathcal{H}}$ from a_T to be equal to the golden entity e_i^* used for constructing q .

Step 2: Citation-based Rubric Judgment. For each fully-identified rubric $r_j^{\mathcal{H}} \in \mathcal{R}_q^{\text{identify}}$, we further check whether $r_j^{\mathcal{H}}$ is grounded on the cited web contents in \mathcal{H} , preventing the agent from fabricating entity names or facts. To achieve this, we first extract cited URLs² from the final response a_T using regex and collect corresponding web contents from \mathcal{H} to form the supporting context $\mathcal{C}^{\mathcal{H}}$:

$$\begin{aligned} url_1, \dots, url_k &= \text{ExtractCitation}(a_T), \\ \mathcal{C}^{\mathcal{H}} &= \text{CollectContent}(\mathcal{H}, url_1, \dots, url_k), \end{aligned} \quad (8)$$

which includes deduplicated search snippets, opened webpage content, and keyword matches. Then we prompt the LLM $\mathcal{M}_{\text{judge}}$ to judge whether each identified rubric is fully supported by $\mathcal{C}^{\mathcal{H}}$:

$$\begin{aligned} \{sp_1, \dots, sp_{m_q}\} &= \mathcal{M}_{\text{judge}}(\mathcal{R}_q^{\text{identify}}, \mathcal{C}^{\mathcal{H}}), \\ \mathcal{R}_q^{\text{support}} &= \{r_j^{\mathcal{H}} \in \mathcal{R}_q^{\text{identify}} \mid sp_j = 1\}, \end{aligned} \quad (9)$$

where $sp_j \in \{0, 1\}$ indicates whether $r_j^{\mathcal{H}}$ is supported.

Step 3: Evidence Connectivity Check. Beyond individual support, we require that supported rubrics form connected evidence chains linked to the predicted answer entity $e_0^{\mathcal{H}}$. This prevents the agent from hacking a rubric by finding entities that satisfy the factual statement but are unrelated to $e_0^{\mathcal{H}}$. Specifically, we construct a bipartite graph:

$$\mathcal{G}^{\mathcal{H}} = \{\mathcal{E}_q^{\mathcal{H}} \cup \mathcal{R}_q^{\text{support}}, E\}, \quad (10)$$

whose nodes are identified entities $\mathcal{E}_q^{\mathcal{H}}$ and supported rubrics $\mathcal{R}_q^{\text{support}}$, with an edge $(e_i^{\mathcal{H}}, r_j^{\mathcal{H}}) \in E$ if $e_i^{\mathcal{H}}$ appears in $r_j^{\mathcal{H}}$, i.e., $e_i^{\mathcal{H}} \in \mathcal{E}_{q,j}^{\mathcal{H}}$. Then we apply a breadth-first search (BFS) starting from $e_0^{\mathcal{H}}$ to determine the set of reachable rubrics $\mathcal{R}^{\text{connect}}$:

$$\mathcal{R}^{\text{connect}} = \{r_j^{\mathcal{H}} \mid r_j^{\mathcal{H}} \text{ is connected to } e_0^{\mathcal{H}} \text{ in } \mathcal{G}^{\mathcal{H}}\} \quad (11)$$

The final rubric reward is given by:

$$R_r^{\mathcal{H}} = \frac{|\mathcal{R}_q^{\text{connect}}|}{|\mathcal{R}_q|}, \quad (12)$$

which measures the proportion of rubrics that are fully identified, citation-supported, and logically connected to the predicted answer.

²We extract at most 20 cited URLs to prevent the agent from hacking the reward by citing a large amount of webpages.

2.3 C-GRPO

Based on CaRR framework, we further introduce **Citation-aware Group Relative Policy Optimization (C-GRPO)**, which combines citation-aware rubric rewards and outcome rewards in GRPO for training robust deep search agents. Specifically, C-GRPO assigns an additional weighted rubric reward to the trajectories whose outcome reward is 1. By doing so, C-GRPO preserves the primary objective of finding the correct answer while encouraging the agent to produce more comprehensive and evidence-grounded reasoning processes. Formally, let $\mathcal{H}_1, \dots, \mathcal{H}_G$ be a group of rollout for a question q , whose ground truth answer is gt . We first obtain the outcome reward $R_o^{\mathcal{H}_i}$ (i.e., whether \mathcal{H}_i finds gt) and the context-aware rubric reward $R_r^{\mathcal{H}_i}$ for each \mathcal{H}_i . Then the mixed reward of \mathcal{H}_i is defined as:

$$R_i = (1 - \alpha) \cdot R_o^{\mathcal{H}_i} + \alpha \cdot R_o^{\mathcal{H}_i} \cdot \hat{R}_r^{\mathcal{H}_i}, \quad (13)$$

where $\alpha \in [0, 1]$ is a hyperparameter balancing outcome and rubric rewards, and we use the normalized rubric rewards

$$\hat{R}_r^{\mathcal{H}_i} = \frac{R_r^{\mathcal{H}_i}}{\max_{j \in \{1, \dots, G\}} R_r^{\mathcal{H}_j}} \quad (14)$$

to stable advantage calculation across different groups. In addition, rollouts with format error or overlength problem (i.e., exceeding token or tool-call limits) are assigned a reward of 0. Finally, the agent policy is optimized by maximizing a multi-turn GRPO objective with token-level loss:

$$\begin{aligned} \mathcal{J}(\theta) = & \mathbb{E}_{\substack{(q, gt) \sim \mathcal{D}, \\ \{\mathcal{H}_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)}} \left[\frac{1}{\sum_{i=1}^G \sum_{j=1}^{|\mathcal{H}_i|} I(\mathcal{H}_{i,j})} \sum_{i=1}^G \sum_{j=1}^{|\mathcal{H}_i|} \right. \\ & \left. I(\mathcal{H}_{i,j}) \min \left(\rho_{i,j} \hat{A}_{i,j}, \text{clip} \left(\rho_{i,j} \frac{1+\epsilon_{\text{high}}}{1-\epsilon_{\text{low}}} \hat{A}_{i,j} \right) \right) \right]. \end{aligned} \quad (15)$$

where $\mathcal{H}_{i,j}$ denotes the j -th token of \mathcal{H}_i ; $\rho_{i,j} = \frac{\pi_\theta(\mathcal{H}_{i,j}|q, \mathcal{H}_{i,1:j-1})}{\pi_{\theta_{\text{old}}}(\mathcal{H}_{i,j}|q, \mathcal{H}_{i,1:j-1})}$ and $\hat{A}_{i,j} = \frac{R_i - \text{mean}(R_k)_{k=1}^G}{\text{std}(R_k)_{k=1}^G}$ are the importance sampling ratio and advantage of $\mathcal{H}_{i,j}$; and $I(\mathcal{H}_{i,j}) \in \{0, 1\}$ indicates whether $\mathcal{H}_{i,j}$ is generated by the LLM itself (i.e., not from observed web content).

3 Experiments

In this section, we conduct RL experiments to show the effectiveness of context-aware rubric rewards and C-GRPO for training deep search agents.

3.1 Experiment Setup

Models and Training Data. We select Qwen3-4B-Thinking-2507 (Team, 2025b) and Qwen3-30B-A3B-Thinking-2507 (Team, 2025b) as our backbone models, covering different model sizes and architectures (dense and MoE). We use DeepDive (Lu et al., 2025), an open-sourced deep search dataset, as our training data. This dataset is automatically synthesized through knowledge graph random walks and entity obfuscation, consisting of 1,016 samples for SFT and 2,234 samples for RL.

Environment Settings We use Serper API (Serper, 2025) for the search tool. The open tool first fetches the webpage using Jina API (Jina.ai, 2025), and then returns the first 10k chars. The find tool is implemented with vanilla string matching.

Baselines. To demonstrate the algorithm enhancement, we compare C-GRPO with two baseline RL algorithms for deep search agents: (1) GRPO (Shao et al., 2024; Jin et al., 2025a), which only uses the outcome rewards and is widely adopted in previous works; (2) E-GRPO (Zhao et al., 2025), which takes normalized entity match rate (i.e., the ratio of golden hidden entities identified during an agent’s reasoning process), as the fine-grained rewards for incorrect rollouts to distinguish “near-miss” samples from complete failures. Besides, we present the reported scores of several state-of-the-art search agents as references, though they may adopt different training data and context lengths from us. We detail them in Appendix B.

Training Details. Our training process includes cold-start SFT and subsequent RL. For cold-start SFT, we first leverage GLM-4.6 (Team, 2025a) to generate 832 high-quality SFT traces through reject sampling on the SFT split of DeepDive dataset. Then we train each model on these traces for 3 epochs with a batch size of 16, a learning rate of 4e-5, and a maximum context length of 128k. For RL, we use all 2,234 QA pairs from the DeepDive RL split. The training configuration includes a rollout size of 16, 8 samples per prompt, a global batch size of 128, a temperature of 1.0, a learning rate of 2e-6, and a maximum context length of 64k tokens. We train each model for 3 epochs. We set the rubric reward weight α to be 0.3, and the effect of different values of α can be found in Sec.. We use DeepSeek-v3.2 (Liu et al., 2025a) as the judge LLM for both outcome rewards and rubric rewards.

Model	Context Budget	BrowseComp	BrowseComp-ZH	xbench-DS	GAIA
<i>Advance Agents with Proprietary Data</i>					
OpenAI-o3	128k	49.7	58.1	66.7	70.5
DeepSeek-v3.1	128k	30.0	49.2	71.2	63.1
Tongyi-DeepResearch	128k	43.4	46.7	75.0	70.9
GLM-4.5	128k	26.4	37.5	70.0	66.0
GLM-4.6	128k	45.1	49.5	-	-
<i>Agents with Open-source Data</i>					
Asearcher-Web-32B	128k	5.2	15.6	42.1	52.8
WebSailor-7B	32k	6.7	14.2	34.3	37.9
WebSailor-32B	32k	10.5	25.5	53.3	51.5
WebExplorer-8B	128k	15.2	32.0	53.7	50.0
DeepDive-9B	128k	6.3	15.1	51.8	-
DeepDive-32B	128k	15.3	29.7	51.8	-
<i>Our Agents</i>					
DeepDive-4B-SFT	64k / 128k	7.7 / 14.1	10.1 / 16.6	34.0 / 44.3	39.5 / 46.0
+ GRPO	64k / 128k	12.9 / 14.7	16.6 / 17.5	41.0 / 41.3	40.5 / 41.1
+ E-GRPO	64k / 128k	11.5 / 14.5	16.5 / 20.2	43.7 / 45.0	42.4 / 42.4
+ C-GRPO (ours)	64k / 128k	13.9 / 17.5	18.2 / 24.7	50.3 / 54.0	48.9 / 50.2
DeepDive-30B-SFT	64k / 128k	12.2 / 20.5	15.8 / 24.7	43.0 / 54.3	46.0 / 50.8
+ GRPO	64k / 128k	16.0 / 18.9	24.1 / 26.1	51.3 / 52.0	51.1 / 51.1
+ E-GRPO	64k / 128k	13.1 / 18.5	17.1 / 24.0	51.7 / 55.7	52.8 / 55.3
+ C-GRPO (ours)	64k / 128k	17.9 / 24.8	26.0 / 33.3	55.3 / 57.7	53.7 / 56.3

Table 1: Overall performance of different agents on four challenging deep search benchmarks.

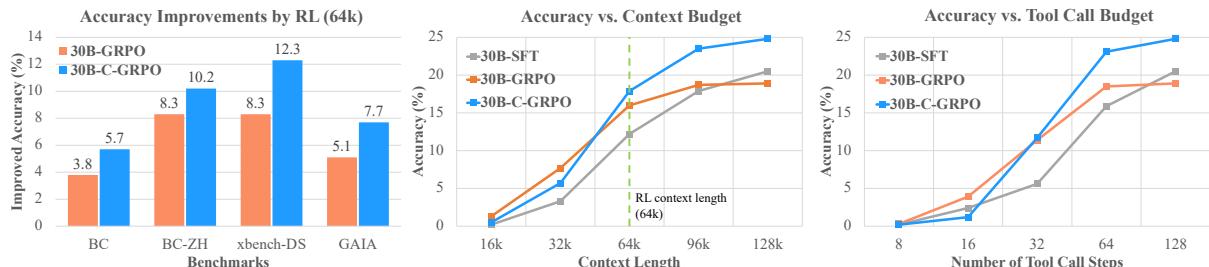


Figure 3: *Left:* Accuracy improvements by GRPO and C-GRPO over SFT models at 64k context length. *Middle* and *Right:* Test-time scaling performance of different models with respect to context budget and tool call budget.

Benchmarks and Evaluation Details. We evaluate the trained agents on four challenging deep search benchmarks, including: BrowseComp (Wei et al., 2025), BrowseComp-ZH (Zhou et al., 2025), xbench-DeepSearch (Xbench-Team, 2025), and the text-only validation subset of GAIA (Mialon et al., 2024). These benchmarks comprehensively assess the essential capabilities for effective deep search in long-horizon information-seeking, multi-step web navigation, complex reasoning, and cross-lingual synthesis. Following their official LLM-as-judge settings, we use GPT-5-Chat (OpenAI, 2025b) to assess whether the agent’s final output matches the ground truth answer. Considering the relatively small dataset size of BrowseComp-ZH, xbench-DeepSearch, and GAIA, we repeat their evaluation 3 times and report the average accuracy. In addition, we evaluate each model at both 64k and 128k

context lengths, where the former corresponds to the context length of RL, and the latter is used to assess the test-time scaling capacities given abundant context budgets.

3.2 Main Result

We present the main experimental results in Table 1. As shown in the table, our proposed C-GRPO significantly outperforms GRPO and E-GRPO baselines on all benchmarks across both 4B and 30B scales. Specifically, with the 64k/128k context budget, C-GRPO achieves an average improvement of 5.1/8.0 for the 4B model and 2.6/6.0 for the 30B model compared to GRPO. Surprisingly, we find that though GRPO with pure outcome reward notably improves the performance of SFT models within the RL context length (i.e., 64k), it may compromise their test-time scaling performance on

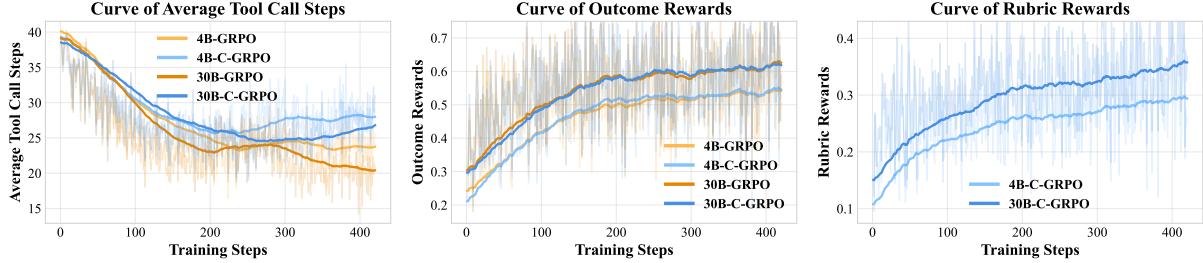


Figure 4: Training dynamics of GRPO and C-GRPO, including the changes of average tool call steps, outcome rewards, and rubric rewards.

Model	$ \mathcal{C}^H $	$ \mathcal{R}_q^{\text{identify}} $	$ \mathcal{R}_q^{\text{support}} $	$ \mathcal{R}_q^{\text{connect}} $	$ \mathcal{R}_q $
DeepDive-30B-SFT	3.8	8.0	6.2	4.5	10.1
+ GRPO	3.5	7.5	5.3	4.0	10.1
+ C-GRPO (ours)	4.3	8.2	6.6	5.2	10.1

Table 2: Comparison of the number of cited webpages and rubric satisfaction on a subset of BrowseComp.

longer context length (i.e., 128k). Our training dynamic analysis and case studies indicate that this compromise stems from the inherent limitation of pure outcome rewards, which leaves room for shortcut exploitation and hallucinations. In contrast, C-GRPO consistently improves the SFT models at both 64k and 128k context lengths, demonstrating the effectiveness of context-aware rubric rewards for training more robust deep search agents. Moreover, our trained models with C-GRPO achieve state-of-the-art performance among agents using open-source data, narrowing the gap with advanced agents that use proprietary data.

3.3 More Analysis

Traning dynamics. We present the training dynamics of GRPO and C-GRPO in Figure 4. As illustrated, the average tool call steps of GRPO and C-GRPO both decline at the beginning of training, where the agents learn to improve their search efficiency to avoid overlength rollouts. As the training progresses, the tool call steps of GRPO continue to decrease after a slight increase, implying that the models fall into a local optimal policy that favors shortcut solutions. Case studies in Appendix D show that the GRPO agent becomes prone to finding an answer based on the last few hops of the question without thoroughly verifying other constraints. While such a policy can yield high outcome rewards within limited context budgets, it sacrifices performance on more difficult questions that require careful verification using longer contexts. Moreover, outcome rewards alone are insuffi-

cient to guide agents out of this local optimum, as they cannot punish the shortcut exploitation behaviors. In contrast, the tool call steps of our C-GRPO keep increasing after the initial decline, suggesting that the models are trying to satisfy more rubrics by gathering more evidence to support and verify their predicted answer, which results in a more robust policy. During the same period, the outcome rewards of C-GRPO even slightly exceed GRPO, further validating that our mixed fine-grained rewards provide a more effective and robust learning signal than pure outcome rewards.

Comprehensiveness and factuality. To assess the impact of different RL algorithms on the comprehensiveness and factuality of agents, we analyze the rubric satisfaction of 30B agents on the evaluation sets. Specifically, we generate rubrics for a subset of BrowseComp in which all agents solve the queries within a 64k context length, and compare the number of their cited webpages and satisfied rubrics. As shown in Table 2, the C-GRPO agent cites more webpages and satisfies more rubrics than the SFT and GRPO baselines, indicating that C-GRPO effectively enhances agent comprehensiveness and factuality by incentivizing more extensive evidence gathering. Conversely, the cited webpages and satisfied rubrics of the GRPO agent are both fewer than the SFT baseline, further validating the shortcut exploitation issue of pure outcome rewards.

Generalize to open-ended deep research tasks. To assess the generalization capabilities of agents trained using C-GRPO and synthetic data in open-ended deep research tasks, we conduct evaluations on DeepResearch Bench (Jin et al., 2025b), where the agents are required to write research reports for PhD-level tasks, and the generated reports are assessed by Gemini-2.5-Pro-preview (Google, 2025) based on the pre-defined rubrics spanning multi-

Model	DeepResearch Bench				
	Overall	Comp.	Insight	Inst.	Read.
Advance Agents with Proprietary Data					
OpenAI-DeepResearch	46.45	46.46	43.73	49.39	47.22
Kimi-Researcher	44.64	44.96	41.97	47.14	45.59
Tongyi-DeepResearch	40.46	39.46	34.44	46.22	44.27
Grok-Deeper-Search	38.22	36.08	30.89	46.59	42.17
Our Agents					
DeepDive-4B-SFT	33.81	29.57	24.23	44.05	41.02
+ GRPO	34.79	31.29	26.79	43.81	41.58
+ E-GRPO	36.59	33.20	28.30	45.58	42.67
+ C-GRPO (ours)	37.51	33.88	30.01	45.72	43.82
DeepDive-30B-SFT	37.51	34.27	28.85	46.77	43.21
+ GRPO	39.30	36.10	31.66	47.65	44.92
+ E-GRPO	36.12	32.31	27.73	45.72	42.33
+ C-GRPO (ours)	41.99	39.75	35.87	48.51	46.63

Table 3: Performance of different agents on DeepResearch Bench across four dimensions, including comprehensiveness (Comp.), insight, instruction following (Inst.), and readability (Read.).

Model	BC	BC-ZH	xbench-DS	GAIA
DeepDive-4B-SFT	14.1	16.6	44.3	46.0
+ C-GRPO ($\alpha=0$)	14.7	17.5	41.3	41.1
+ C-GRPO ($\alpha=0.1$)	13.0	18.0	46.0	46.3
+ C-GRPO ($\alpha=0.3$)	17.5	24.7	54.0	50.2
+ C-GRPO ($\alpha=0.5$)	17.0	20.8	49.3	42.4

Table 4: Performance of C-GRPO with different α .

Model	BC	BC-ZH	xbench-DS	GAIA
DeepDive-4B-C-GRPO	17.5	24.7	54.0	50.2
w/o Hidden entity identification	16.5	23.2	50.7	46.6
w/o Evidence connectivity check	15.1	20.8	47.7	44.0
w/ Rubric rewards for all rollouts	13.3	14.0	40.3	40.8

Table 5: Performance of C-GRPO (1) without hidden entity identification; (2) without evidence connect check; (3) that adds weighted rubric rewards for all rollouts.

ple dimensions. As shown in Table 3, C-GRPO consistently surpasses other RL algorithms and yields substantial improvements over SFT models in all dimensions. Moreover, the 30B model trained with C-GRPO even outperforms several advanced agents using proprietary data, demonstrating the strong generalization abilities of our approach.

3.4 Ablation Studies

In this section, we conduct ablation studies using the 4B model to demonstrate the effect of each component in the CaRR framework and C-GRPO.

Effect of rubric reward weight. To illustrate the effect of the rubric reward weight α in C-GRPO, we train the 4B model with different α values, ranging from 0 (which is just GRPO) to 0.5. As shown in Table 4, the overall performance grad-

ually improves as α increases from 0, peaking at 0.3. This demonstrates the benefit of incorporating context-aware rubric rewards in RL. However, the performance begins to decrease as α becomes larger, suggesting that the model is distracted from the primary goal of finding a correct final answer. Therefore, it is important to use a moderate α value to balance the two reward components to obtain the optimal policy.

Effect of hidden entity identification. To show the effect of hidden entity identification in CaRR, we remove this step and let the judge LLM directly select the supported rubrics based on the model response and cited web contents, without considering whether each rubric is fully identified. As shown in Table 5, this ablation leads to a clear performance drop for C-GRPO, suggesting that enforcing a stricter reward process via hidden entity identification enhances the effectiveness of RL.

Effect of evidence connectivity check. We show the effect of the evidence connectivity check in CaRR by eliminating this step and instead setting the rubric reward to the fraction of supported rubrics, i.e., $R_r^{\mathcal{H}} = \frac{\mathcal{R}_q^{\text{support}}}{|\mathcal{R}_q|}$. The results in Table 5 show a substantial decline in performance without the connectivity check, since the agents learns to hack rubrics by finding entities that satisfy isolated factual statements but are unrelated to the final answer.

Adding rubric rewards for all rollouts. According to Equation 13, we only add weighted context-aware rubric rewards for correct rollouts whose outcome reward is 1. If we add the rubric rewards for all rollouts, the advantage of some incorrect rollouts will receive positive advantages when there are few correct rollouts or many overlength rollouts in a group, which frequently happens at the beginning of RL. As a result, the model will be incorrectly optimized and perform badly, as shown in Table 5.

4 Related Works

RL for Deep Search Agents. Recently, RL has emerged as a critical technique for enhancing deep search agents (OpenAI, 2025a; Jin et al., 2025a). Existing works can be broadly divided into two categories. The first category focuses on complex QA data synthesis and infrastructure design to support RL training (Gao et al., 2025; Wu et al., 2025a; Li et al., 2025b; Lu et al., 2025; Liu et al., 2025b; Li

et al., 2025a; Team et al., 2025), and the second category focuses on improving RL algorithms to better fit multi-turn agentic settings (Feng et al., 2025; Dong et al., 2025c,a). Nonetheless, these works typically rely solely on outcome rewards, with limited attention devoted to addressing the their limitations. E-GRPO (Zhao et al., 2025) proposes to use the entity match rate as the fine-grained rewards for incorrect rollouts to distinguish “near-miss” samples from complete failures. However, it relies on gold annotations for intermediate hidden entities, and we also observed that applying fine-grained rewards for incorrect rollouts may mislead the RL optimization (see Sec. 3.4).

Aligning LLMs with Rubric Rewards. Recently, a series of works have explored the use of rubrics (Arora et al., 2025; Asai et al., 2024) in aligning LLMs for complex instruction following (Lambert et al., 2024; Dong et al., 2025b; Peng et al., 2025) and long-form generation tasks (Gunjal et al., 2025; Shao et al., 2025a), where traditional reward models fail to provide reliable supervision signals. Specifically, they equip each training instance with a list of verifiable rubrics, and the reward of a model response is given by the ratio of its satisfied rubrics. Some works also explore evolving rubrics during training by contrasting multiple model rollouts (Shao et al., 2025a; Rezaei et al., 2025; Wu et al., 2025b). In this work, we show that rubric rewards can be utilized to supervise agents’ reasoning processes, serving as an effective auxiliary of traditional outcome rewards.

5 Conclusion

In this work, we propose CaRR, a novel framework that provides fine-grained rewards for deep search agents, taking into account reasoning comprehensiveness, factual grounding, and evidence connectivity. We further introduce C-GRPO, which combines CaRR and outcome rewards in RL for training robust deep search agents. Our extensive experiments demonstrate that C-GRPO achieves significant improvement over GRPO in both deep search benchmarks and open-ended research tasks.

6 Limitations

As described in Sec. 2, our rubric generation relies on the compositional structure of synthetic multi-hop questions, and may not be able to be directly adapted to open-ended QA training where some

requirements are not explicitly stated in the question. Nonetheless, both our work and previous works have shown that the synthetic, short-form question answering is an effective proxy for open-ended deep research tasks since they share the core requirement for long-horizon information-seeking capacity. Moreover, the improvement in reasoning comprehensiveness and factual grounding brought by our context-aware rubric rewards also benefits the model’s performance on open-ended deep research tasks, as demonstrated in Sec. 3.3.

7 Ethical Considerations

All the models and datasets used in this work are publicly published with permissible licenses.

References

- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#). *CoRR*, abs/2505.08775.
- Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D’Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. [Openscholar: Synthesizing scientific literature with retrieval-augmented lms](#). *CoRR*, abs/2411.14199.
- DeepSeek. 2025. [Deepseek-v3.1 release](#).
- Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2025a. [Agentic entropy-balanced policy optimization](#). *CoRR*, abs/2510.14545.
- Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. 2025b. [Self-play with execution feedback: Improving instruction-following capabilities of large language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazheng Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2025c. [Agentic reinforced policy optimization](#). *CoRR*, abs/2507.19849.

- Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025. [Group-in-group policy optimization for LLM agent training](#). *CoRR*, abs/2505.10978.
- Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. 2025. [Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous RL](#). *CoRR*, abs/2508.07976.
- Google. 2025. [Gemini 2.5 pro preview: even better coding performance](#).
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. 2025. [Rubrics as rewards: Reinforcement learning beyond verifiable domains](#). *CoRR*, abs/2507.17746.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025a. [Search-r1: Training llms to reason and leverage search engines with reinforcement learning](#). *CoRR*, abs/2503.09516.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025b. [Search-r1: Training llms to reason and leverage search engines with reinforcement learning](#). *CoRR*, abs/2503.09516.
- Jina.ai. 2025. [Jina](#).
- Kimi. 2025. [Kimi-researcher: End-to-end rl training for emerging agentic capabilities](#).
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. [Tülu 3: Pushing frontiers in open language model post-training](#). *CoRR*, abs/2411.15124.
- Kuan Li, Zhongwang Zhang, Hufeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litu Ou, Dingchu Zhang, Xixi Wu, Jialong Wu, Xinyu Wang, Zile Qiao, Zhen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025a. [Websailor-v2: Bridging the chasm to proprietary agents via synthetic data and scalable reinforcement learning](#). *CoRR*, abs/2509.13305.
- Kuan Li, Zhongwang Zhang, Hufeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025b. [Websailor: Navigating super-human reasoning for web agent](#). *CoRR*, abs/2507.02592.
- Aixin Liu, Aoxue Mei, Bangcui Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. [Deepseek-v3. 2: Pushing the frontier of open large language models](#). *arXiv preprint arXiv:2512.02556*.
- Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhui Chen, Pengyu Zhao, and Junxian He. 2025b. [Webexplorer: Explore and evolve for training long-horizon web agents](#). *CoRR*, abs/2509.06501.
- Rui Lu, Zhenyu Hou, Zihan Wang, Hanchen Zhang, Xiao Liu, Yujiang Li, Shi Feng, Jie Tang, and Yuxiao Dong. 2025. [Deepdive: Advancing deep search agents with knowledge graphs and multi-turn RL](#). *CoRR*, abs/2509.10446.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. [GAIA: a benchmark for general AI assistants](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- OpenAI. 2025a. [Deep research system card](#).
- OpenAI. 2025b. [Introducing gpt-5](#).
- OpenAI. 2025c. [Introducing o3 and o4-mini](#).
- Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2025. [Verif: Verification engineering for reinforcement learning in instruction following](#). *CoRR*, abs/2506.09942.
- MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyürek. 2025. [Online rubrics elicitation from pairwise comparisons](#). *CoRR*, abs/2510.07284.
- Serper. 2025. [Serper: Google search api](#).
- Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, and 1 others. 2025a. [Dr tulu: Reinforcement learning with evolving rubrics for deep research](#). *arXiv preprint arXiv:2511.19399*.
- Zhihong Shao, Yuxiang Luo, Chengda Lu, ZZ Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. 2025b. [Deepseekmath-v2: Towards self-verifiable mathematical reasoning](#). *arXiv preprint arXiv:2511.22570*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *CoRR*, abs/2402.03300.
- GLM Team. 2025a. [GLM-4.5: agentic, reasoning, and coding \(ARC\) foundation models](#). *CoRR*, abs/2508.06471.
- Qwen Team. 2025b. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025. Tongyi deepreresearch technical report. *arXiv preprint arXiv:2510.24701*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, 18(6):186345.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecamp: A simple yet challenging benchmark for browsing agents. *CoRR*, abs/2504.12516.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025a. Webdancer: Towards autonomous information seeking agency. *CoRR*, abs/2505.22648.

Mian Wu, Gavin Zhang, Sewon Min, Sergey Levine, and Aviral Kumar. 2025b. RLAC: reinforcement learning with adversarial critic for free-form generation tasks. *CoRR*, abs/2511.01758.

xAI Team. 2025. Grok agents: Combining reasoning and tool use. <https://x.ai/news/grok-3#grok-agents-combining-reasoning-and-tool-use>.

Xbench-Team. 2025. Xbench-deepsearch.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yida Zhao, Kuan Li, Xixi Wu, Liwen Zhang, Dingchu Zhang, Baixuan Li, Maojia Song, Zhuo Chen, Chenxi Wang, Xinyu Wang, Kewei Tu, Pengjun Xie, Jingren Zhou, and Yong Jiang. 2025. Repurposing synthetic data for fine-grained search agent supervision. *CoRR*, abs/2510.24694.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. 2025. Browsecamp-zh: Benchmarking web browsing ability of large language models in chinese. *CoRR*, abs/2504.19314.

A Trajectory Format

We show our tool descriptions and trajectory format in Figure 5 and 6, respectively.

B Details of Referred Deep Search Agents

For deep search benchmarks, we present the scores of OpenAI o3 (OpenAI, 2025c), DeepSeek-v3.1 (DeepSeek, 2025), Tongyi-DeepResearch (Team et al., 2025), GLM-4.5 (Team, 2025a), GLM-4.6 (Team, 2025a), Aseacher (Gao et al., 2025), WebSailor (Li et al., 2025b), WebExplorer (Liu et al., 2025b), and DeepDive (Lu et al., 2025) from their official reports or previous papers. For DeepResearch Bench, we present the scores of OpenAI-DeepDeesearch (OpenAI, 2025a), Kimi-Researcher (Kimi, 2025), Tongyi-DeepResearch (Team et al., 2025), and Grok-Deeper-Search (xAI Team, 2025) from the official leaderboard (Jin et al., 2025b).

C Human Verification for LLM Judge

To assess the reliability of the judge LLM in identifying hidden entities and applying citation-based rubric evaluations within the CaRR framework, we conducted a manual review of its judgments across 10 DeepDive-30B-SFT trajectories, covering 128 hidden entities and 164 rubrics. Using human assessments as the gold standard, the judge LLM achieved accuracies of 97.7% for hidden entity identification and 95.1% for citation-based rubric evaluation, indicating strong reliability.

D Case Studies

To highlight the qualitative differences of GRPO and C-GRPO, we compare trajectories produced by DeepDive-30B-GRPO and DeepDive-30B-C-GRPO for the same queries in both the training set (DeepDive) and the evaluation set (BrowseComp). We present only the final turn of each trajectory, which is sufficient to demonstrate the key distinctions. As shown in Case 1 (Figure 7), the GRPO agent tends to infer the answer based on the last several hops in the question, without carefully checking other constraints. It often guesses the identities of entities in the head part of the question (marked in red) without further verification. Such an unrobust policy is prone to failure on more challenging questions that demand thorough validation, as illustrated in Case 4 (Figure 9). In contrast, the C-GRPO agent, as indicated by the

green highlights in Case 2 (Figure 8) and Case 4 (Figure 10), continues to gather evidence until it can confirm that every constraint in the query is satisfied. It further ensures that each statement in its response is supported by corresponding citations.

E Prompts

We show our used prompts in Figure 11, 12, 13, and 14.

Tool Descriptions

browser.search

Description: Search in browser

Parameters:

- query: string, Search query
- num: integer, Number of results to return (default 10)

browser.open

Description: Open browser link

Parameters:

- id: integer or string, ID or URL of the link to open

browser.find

Description: Find pattern in the opened browser content

Parameters:

- pattern: string, Pattern to find

Figure 5: Description of search, open, and find tool.

Trajectory Format

```
<|im_start|>user
question <|im_end|>
<|im_start|>assistant
<think>
thinking process
</think>
<tool_call>
{"name": "tool name", "arguments": {"parameter name": parameter value, "another parameter name": another
parameter value, ...}}
</tool_call><|im_end|>
<|im_start|>user
<tool_response>
tool response
</tool_response><|im_end|>
(more thinking processes, tool calls, and tool responses)
<|im_start|>assistant
<think>
thinking process
</think>
## Explanation with Citations
explanation with in-line citations
## Exact Answer
final answer <|im_end|>
```

Figure 6: Format of deep search agent trajectory.

Case 1 on DeepDive: Shortcut Solution by DeepDive-30B-GRPO

Question: Begin with an entity that raises money for three health organizations. One of these, a medical establishment whose name suggests a 'God's guesthouse', later occupied the site of a former religious college in a city famed for its stone buildings after a major blaze and which once served as a country's capital. This city's public transit serves a royal military school and connects to an intercity bus depot on a boulevard named for a historical figure. This depot provides access to a major numbered highway, a key east-west artery. A scenic section of this artery, running alongside a large international river, later became a distinct roadway. To what entity was the oversight of this distinct roadway transferred in the early 1970s?

Answer: St. Lawrence Parks Commission

DeepDive-30B-GRPO:

Round 1

...

Round 21

```
<|im_start|>assistant  
<think>
```

Scrolling near the part referencing the scenic highway. Already have. Now, we also need a citation for the transfer to the St. Lawrence Parks Commission. That is in the same article. Now we need to reference the "city famed for its stone buildings after a major blaze and once served as a country's capital". The city likely is "Ottawa"? Or maybe "St. Paul"? ... But not. But we can provide a citation for the transfer of the road, which is the core. Thus for explanation we can mention that the road is part of Highway 401 and later became the Thousand Islands Parkway ... Now produce answer.

```
</think>
```

Explanation with Citations

The major east–west artery is Ontario's Highway 401, which carries traffic across the province [https://en.wikipedia.org/wiki/Highway_401]. A scenic section of this highway that runs alongside the international St. Lawrence River was originally designated Highway 2S and later became the Thousand Islands Parkway [https://en.wikipedia.org/wiki/Thousand_Islands_Parkway]. After the Parkway was decommissioned on 8 September 1970, jurisdiction over the road was transferred to the St. Lawrence Parks Commission [https://en.wikipedia.org/wiki/Thousand_Islands_Parkway].

Thus, the early-1970s transfer of oversight was to the St. Lawrence Parks Commission.

Exact Answer

St. Lawrence Parks Commission <|im_end|> ✓

Figure 7: A case from the DeepDive dataset where DeepDive-30B-GRPO solves the question via shortcut solution.

Case 2 on DeepDive: Fully Solved by DeepDive-30B-C-GRPO

Question: Begin with an entity that raises money for three health organizations. One of these, a medical establishment whose name suggests a 'God's guesthouse', later occupied the site of a former religious college in a city famed for its stone buildings after a major blaze and which once served as a country's capital. This city's public transit serves a royal military school and connects to an intercity bus depot on a boulevard named for a historical figure. This depot provides access to a major numbered highway, a key east-west artery. A scenic section of this artery, running alongside a large international river, later became a distinct roadway. To what entity was the oversight of this distinct roadway transferred in the early 1970s?

Answer: St. Lawrence Parks Commission

DeepDive-30B-C-GRPO:

Round 1

...

Round 41

```
<|im_start|>assistant  
<think>
```

Thus, the entity is University Hospitals Kingston Foundation (UHKF). It raises money for three health organizations: Kingston General Hospital, Hotel-Dieu Hospital, and Providence Care. Now we have all needed steps. Now we must answer: "To what entity was the oversight of this distinct roadway transferred in the early 1970s?" The answer: St. Lawrence Parks Commission (also known as the St. Lawrence Parks Commission). The article states that "On September 8, 1970, the DHO transferred jurisdiction over the parkway to the St. Lawrence Parks Commission." This is early 1970s. Thus final answer: The St. Lawrence Parks Commission. Now we need to produce a final response with Explanation and citations. We'll need to cite: UHKF raising money for three health organizations: from CanadaHelps page [[13]]. Hotel-Dieu in Kingston and former Regiopolis College: from Kingston Museums blog [[1]]. Kingston city: stone buildings, capital, major fire: from Wikipedia pages: Kingston, Ontario [[12]] lines about nickname "Limestone City," capital, fire; also the 1840 fire. Public transit serves Royal Military College: from Kingston Transit route pdf and RMC Wikipedia. Intercity bus depot on John Counter Boulevard: from Kingston Bus Terminal page [[9]]. Major highway 401: from Highway 401 page [[11]] and Kingston Bus Terminal page says near Highway 401. Scenic section: Thousand Islands Parkway page [[15]]. Transfer: same article lines 3–4. Thus we will craft final answer. Now ensure all factual claims have citations. Ok.

```
</think>
```

Explanation with Citations

1. **Fundraising entity** – The University Hospitals Kingston Foundation (UHKF) is a charitable organization that raises money for three health organizations: Kingston General Hospital, Hotel-Dieu Hospital, and Providence Care [<https://www.canadahelps.org/en/charities/university-hospitals-kingston-foundation>].
2. **Medical establishment with a "God's guesthouse" name** – Among the hospitals it supports is the former Hotel-Dieu Hospital in Kingston, Ontario. Its name translates from French as "House of God," i.e., a "God's guesthouse" [<https://museumskingston.ca/the-founding-of-the-hotel-dieu-hospital-175-years-ago>].
3. **Location in a stone-built city that later held a former religious college** – In 1892 the Hotel-Dieu moved into the former Regiopolis College building on Sydenham Street in Kingston [<https://museumskingston.ca/the-founding-of-the-hotel-dieu-hospital-175-years-ago>]. Kingston is nicknamed the "Limestone City" because many of its historic buildings are constructed of local limestone [https://en.wikipedia.org/wiki/Kingston,_Ontario]. It suffered a major fire in 1840 after a dock explosion, which destroyed much of the downtown [https://en.wikipedia.org/wiki/Kingston,_Ontario]. Kingston also served as the first capital of the United Province of Canada (1841–1844) [https://en.wikipedia.org/wiki/Kingston,_Ontario].
4. **Public transit serving a royal military school** – Kingston's public-transit system (Kingston Transit) operates route 601/602 that connects the downtown core directly to the Royal Military College of Canada in Kingston [https://www.kingstontransit.ca/media/acwldlbl/kingstontransit_route601-602_may2025-fin.pdf]; the college is a degree-granting military academy [https://en.wikipedia.org/wiki/Royal_Military_College_of_Canada].
5. **Intercity bus depot on a boulevard named for a historical figure** – The Kingston Bus Terminal is located on John Counter Boulevard; John Counter was a notable Canadian politician, making the boulevard named after a historical figure [https://en.wikipedia.org/wiki/Kingston_Bus_Terminal]. The terminal provides access to Highway 401, the principal east-west artery of southern Ontario [https://en.wikipedia.org/wiki/Kingston_Bus_Terminal, https://en.wikipedia.org/wiki/Ontario_Highway_401].
6. **Scenic section alongside a large international river that later became a distinct roadway** – East of Gananoque, Highway 401 runs along the north shore of the St. Lawrence River. The picturesque stretch known as the Thousand Islands Parkway was originally part of Highway 401 but was later separated and designated as a scenic parkway (Highway 2S) before becoming a distinct roadway [https://en.wikipedia.org/wiki/Thousand_Islands_Parkway].
7. **Transfer of oversight in the early 1970s** – On 8 September 1970, the jurisdiction over the Thousand Islands Parkway was transferred from the Ontario Department of Highways to the St. Lawrence Parks Commission, which continues to maintain it today [https://en.wikipedia.org/wiki/Thousand_Islands_Parkway].

Thus, the entity that assumed oversight of the scenic roadway in the early 1970s is the **St. Lawrence Parks Commission**.

Exact Answer

St. Lawrence Parks Commission <|im_end|> ✓

Figure 8: A case from the DeepDive dataset where DeepDive-30B-C-GRPO completely solves the question via rigorous verification.

Case 3 on BrowseComp: Failed by DeepDive-30B-GRPO

👤 **Question:** I'm looking for information about a spy who spoke more than one language and was mentioned in a 2023 review of a book about the period in which they were active. They were married to another spy who had operated on different continents and been detected twice: firstly, when they were arrested alongside two Americans by police in a European country who had foiled their network, and secondly, when they inadvertently blew their cover in Asia to a group of foreign officers. They also worked as the handler for a spy whose grandchild wrote about them in a 2014 article about the publication of an intelligence service's archived documents. These two spies had a child together who would go on to be arrested as part of a dissident group, pleading not guilty when charged and serving shorter than their original prison sentence. The first spy I mentioned once helped a journalist with research for a novel. What was the nationality of this journalist?

Answer: Australian

🤖 **DeepDive-30B-GRPO:**

Round 1

...

Round 37

```
<|im_start|>assistant  
<think>
```

Thus David Ignatius is American. Now we need to find if a spy helped him research a novel. Possibly "the CIA officer 'Robert Gates' or 'James Woolsey' but not. But perhaps the first spy is "the 'CIA's 'Philip'. But maybe it's about "the journalist 'David Ignatius' being helped by a spy named 'John K. Cooley'. Given difficulty, maybe the first spy is "the 'British spy 'Sir Richard Dearlove'... Ok I'll answer "American". Thus final answer: The journalist was American. Cite Wikipedia page showing David Ignatius is American journalist and novelist. [[1]]. But the question may refer to a different journalist but nationality still American. So answer "American". Now produce final answer.

```
</think>
```

Explanation with Citations

The journalist in question is **David Ignatius**, an American columnist and author of several espionage novels. He is described in his biographical entry as "an American journalist and novelist" [https://en.wikipedia.org/wiki/David_Ignatius].

Ignatius has credited former intelligence officials with helping him research his novels; for example, in interviews he has mentioned that CIA officers provided background for his thriller *The Director* [<https://www.npr.org/2020/05/04/849927443/in-the-paladin-ignatius-navigates-the-line-between-truth-fiction>].

Because the question asks for the nationality of the journalist who received research assistance from a spy, the answer is that the journalist is **American**.

Exact Answer

American <|im_end|> X

Figure 9: A case from the BrowseComp where DeepDive-30B-GRPO fails due to shortcut exploitation.

Case 4 on BrowseComp: Fully Solved by DeepDive-30B-C-GRPO

Question: I'm looking for information about a spy who spoke more than one language and was mentioned in a 2023 review of a book about the period in which they were active. They were married to another spy who had operated on different continents and been detected twice: firstly, when they were arrested alongside two Americans by police in a European country who had foiled their network, and secondly, when they inadvertently blew their cover in Asia to a group of foreign officers. They also worked as the handler for a spy whose grandchild wrote about them in a 2014 article about the publication of an intelligence service's archived documents. These two spies had a child together who would go on to be arrested as part of a dissident group, pleading not guilty when charged and serving shorter than their original prison sentence. The first spy I mentioned once helped a journalist with research for a novel. What was the nationality of this journalist?

Answer: Australian

DeepDive-30B-C-GRPO:

Round 1

...

Round 54

<|im_start|>assistant

<think>

Thus there is a source confirming that Ulanovsky's cover was blown in Shanghai (in Asia). Great! That matches the second detection: "inadvertently blew their cover in Asia to a group of foreign officers." Actually the article says "cover blown" to a former acquaintance; but we can cite that. Thus all pieces are covered. Now we have enough sources. Now we must confirm that Nadezhda Ulanovskaya is multilingual: Wikipedia includes "studied at the Institute of Foreign Languages" and "translator". That is enough. Now we need to confirm the 2023 review: The Kirkus review mentions Nadya Ulanovskaya; that qualifies. ... Now confirm Maya Ulanovskaya's details: Wikipedia page includes her dissident status, arrest, sentence; we can cite. Now confirm Godfrey Blunden: History Today review includes "helped an Australian correspondent, Godfrey Blunden." Good. Now we can answer: Nationality: Australian. Now we need to provide explanation with citations. Now we need to produce final answer. Now we need to gather all sources for references. List: 1. Nadezhda Ulanovskaya Wikipedia. 2. Kirkus review of The Red Hotel. 3. Alexander Ulanovsky fandom. 4. The Spy in Shanghai article (Michael Batson) for cover blown. 5. MI5 Papers on Hiss-Chambers case article (David Chambers). 6. Maya Ulanovskaya Wikipedia. 7. History Today review (Alan Philps) for journalist. 8. Might also need to cite that Nadezhda was spouse of Alexander Ulanovsky – from Wikipedia. Now produce final answer. Now we need to ensure each claim in explanation has inline citation. Now produce final answer.

</think>

Explanation with Citations

The puzzle describes a network of Soviet-era spies whose identities and relationships can be traced through public records and published accounts.

1. **First spy (the multilingual spouse)** – Nadezhda (Nadya) Ulanovskaya was a Soviet GRU officer, translator, and English teacher, i.e., a multilingual operative [https://en.wikipedia.org/wiki/Nadezhda_Ulanovskaya]. She was featured in a 2023 book review of *The Red Hotel* (a history of the Metropol Hotel in Moscow, 1941) that explicitly mentions her [<https://www.kirkusreviews.com/book-reviews/alan-philips/the-red-hotel-moscow>].

2. **Second spy (the husband, operated on multiple continents, detected twice)** – Alexander Ulanovsky, Nadezhda's husband, was a GRU *rezident* who worked in Europe (Copenhagen, Berlin), the Americas (United States, Shanghai), and Asia [https://military-history.fandom.com/wiki/Alexander_Ulanovsky].

• *First detection (Copenhagen, 1935)* – Danish police arrested Ulanovsky together with two Americans, Leon Josephson and George Mink, foiling the Soviet spy ring [https://military-history.fandom.com/wiki/Alexander_Ulanovsky].

• *Second detection (Shanghai, 1930)* – Ulanovsky's cover was blown in Shanghai when he ran into an old acquaintance who recognized him, forcing his escape [<https://michaelbatson.com/blog/237-the-spy-in-shanghai-18-november-2020>].

3. **Handler for a spy whose grandchild wrote a 2014 article about an intelligence-service archive** – Ulanovsky was the chief illegal *rezident* who handled Whittaker Chambers, the American former spy [https://military-history.fandom.com/wiki/Alexander_Ulanovsky]. The grandchild of Whittaker Chambers, David Chambers, authored a 2014 article on the MI5 release of the Hiss-Chambers case documents [<https://whittakerchambers.org/2014/03/06/mi5-papers-on-hiss-chambers-case>].

4. **Child of the couple (dissident, arrested, pleaded not guilty, shorter sentence)** – Maya Ulanovskaya, their daughter, was a Soviet dissident. She was arrested in 1951, sentenced to 25 years, but the term was reduced to five years and she was released in 1956 [https://en.wikipedia.org/wiki/Maya_Ulanovskaya].

5. **First spy helped a journalist with research for a novel** – In *The Red Hotel*, Alan Philps notes that Nadezhda Ulanovskaya assisted Australian correspondent Godfrey Blunden in researching his novel, for which she was later imprisoned [<https://www.historytoday.com/archive/review/red-hotel-alan-philips-review>].

Conclusion – The journalist who received help from the first spy (Nadezhda Ulanovskaya) was Godfrey Blunden, an Australian correspondent. Therefore, the nationality of the journalist is **Australian**.

Exact Answer

Australian <|im_end|> ✓

Figure 10: A case from BrowseComp where DeepDive-30B-C-GRPO completely solves the question via rigorous verification.

Prompt for Outcome Rewards

Judge whether the following [response] to [question] is correct or not based on the precise and unambiguous [correct_answer] below.

[question]:

{question}

[response]:

{response}

Your judgement must be in the format and criteria specified below:

extracted_final_answer:

The final exact answer extracted from the [response]. Put the extracted answer as None if there is no exact, final answer to extract from the response.

[correct_answer]:

{correct_answer}

reasoning:

Explain why the **extracted_final_answer** is correct or incorrect based on [correct_answer], focusing only on whether there are meaningful differences between [correct_answer] and the **extracted_final_answer**. Do not comment on any background to the problem. Do not attempt to solve the problem. Do not argue for any answer different from [correct_answer]. Focus only on whether the answers match.

correct:

Answer yes if the **extracted_final_answer** matches the [correct_answer] given above, or is within a small margin of error for numerical problems. Answer no otherwise, including cases of inconsistency, ambiguity, non-equivalency, or incorrectness.

Figure 11: Prompt for outcome rewards.

Prompt for Rubric Initialization

You will receive a **complex multi-hop question**. Let $\langle E0 \rangle$ be the final answer entity to the question.

Your task is to break down the question into a list of constraints that $\langle E0 \rangle$ should satisfy.

Requirements for the constraints:

1. Each constraint must be a **single-hop factual statement**, where the intermediate entities should be denoted as $\langle E1 \rangle$, $\langle E2 \rangle$, $\langle E3 \rangle$, and so on.
2. Each constraint must contain **at least one entity**.
3. Each statement must be **clear, coherent, and grammatically correct**.
4. **Do not attempt to infer or guess** the actual identities of any entities.

Your output must follow this format exactly:

```
[Begin of Constraints]
C1. {{constraint 1}}
C2. {{constraint 2}}
C3. {{constraint 3}}
...
[End of Constraints]
```

Here is an example:

[Begin of Question]

Start with a rural settlement in an Asian province, birthplace of a notable female activist. An educational facility in the same administrative division, which transitioned to a secondary school level during a period of widespread international conflict, educated a male activist. This male activist, an early prominent member of a political movement, later entered into matrimony with the aforementioned female activist while both were in a European nation. The female activist, herself an early adherent to this movement and a leader in its female-focused section, eventually engaged in information dissemination work for associated labor bodies in a significant urban center. This political movement's central institution for ideological guidance and information control, founded in the fifth month of a year in the early 1920s, initially included five specific operational units. What was the name of the unit dedicated to managing current affairs and media reports?

[End of Question]

[Begin of Constraints]

- C1. $\langle E1 \rangle$ is a rural settlement in an Asian province, $\langle E2 \rangle$.
- C2. $\langle E1 \rangle$ is the birthplace of a notable female activist $\langle E3 \rangle$.
- C3. $\langle E4 \rangle$ is an educational facility in $\langle E2 \rangle$.
- C4. $\langle E4 \rangle$ transitioned to a secondary school level during a period of widespread international conflict $\langle E5 \rangle$.
- C5. $\langle E4 \rangle$ educated a male activist $\langle E6 \rangle$.
- C6. $\langle E6 \rangle$ was an early prominent member of a political movement $\langle E7 \rangle$.
- C7. $\langle E6 \rangle$ married $\langle E3 \rangle$ while both were in a European nation $\langle E8 \rangle$.
- C8. $\langle E3 \rangle$ was an early adherent to $\langle E7 \rangle$.
- C9. $\langle E3 \rangle$ was a leader in the female-focused section $\langle E9 \rangle$ of $\langle E7 \rangle$.
- C10. $\langle E3 \rangle$ eventually worked in information dissemination for labor bodies $\langle E10 \rangle$ in a significant urban center $\langle E11 \rangle$.
- C11. $\langle E7 \rangle$ had a central institution $\langle E12 \rangle$ for ideological guidance and information control.
- C12. $\langle E12 \rangle$ was founded in the fifth month of the year $\langle E13 \rangle$ in the early 1920s.
- C13. $\langle E12 \rangle$ initially included five specific operational units, including $\langle E0 \rangle$.
- C14. $\langle E0 \rangle$ was a unit dedicated to managing current affairs and media reports.

[End of Constraints]

Now, list the constraints for the following question.

[Begin of Question]

{question}

[End of Question]

Figure 12: Prompt for rubric initialization in CaRR.

Prompt for Entity Identification

Task Description

You will receive the following inputs:

1. A **complex multi-hop question**.
2. A **list of single-hop constraints**, decomposed from the original question.
 - The final answer entity is labeled as <E0>.
 - Intermediate entities are labeled as <E1>, <E2>, <E3>, and so on.
3. An **AI assistant's response** to the multi-hop question.

Your Task

Extract the **explicitly stated real identities** of <E0>, <E1>, <E2>, . . . from the assistant's response. Provide an analysis first, then return a JSON object with one key per entity label.

Output Format

```
## Analysis
{{Explain, for each entity label, whether the assistant's response clearly
and explicitly provides its real identity. State the actual name or value
if it is explicitly mentioned; otherwise, indicate that the identity is
not clearly stated.}}

## Final JSON-format Summary
```json
{
 "E0": {{actual identity from assistant's response or null}},
 "E1": {{actual identity from assistant's response or null}},
 "E2": {{actual identity from assistant's response or null}},
 ...
}
```

```

Important Rules

- Only use information that is **explicitly and unambiguously stated** in the assistant's response.
- **Do not infer, guess, or deduce** entity identities beyond what is explicitly provided.
- If the assistant's response does **not clearly identify** an entity, set its value to null.
- Follow the output format exactly.

[Begin of Question]

{question}

[End of Question]

[Begin of Constraints]

{constraints}

[End of Constraints]

[Begin of Assistant's Response]

{response}

[End of Assistant's Response]

Figure 13: Prompt for entity identification in CaRR.

Prompt for Citation-based Rubric Judgment

You will receive:

1. The contents of several webpages.
2. Several single-hop factual statements: S1, S2, ..., Sn.

Your task is to determine whether each statement is **fully supported** by the provided webpage contents.

For each statement:

- Find the **exact evidence** from the webpage contents that supports or contradicts it.
- Explain clearly why the statement **is** or **is not** fully supported, citing relevant parts of the provided text.
- List the **URLs of webpages** where the supporting evidence was found.
- Conclude with a judgment: **Fully Supported: yes** or **Fully Supported: no**.

At the end, summarize your results in a JSON object mapping each statement label (S1, S2, ...) to a boolean value (**true** for fully supported, **false** for not fully supported).

Output Format:

```
## Supportness Analysis of S1
Explanation: {{Clearly explain why S1 is or is not fully supported
according to the given webpage contents}}
Evidence URLs: {{List of URLs containing evidence used in your explanation}}
Fully Supported: {{yes/no}}

## Supportness Analysis of S2
Explanation: {{Clearly explain why S2 is or is not fully supported
according to the given webpage contents}}
Evidence URLs: {{List of URLs containing evidence used in your explanation}}
Fully Supported: {{yes/no}}
```

...

```
## Final JSON-format Summary
```json
{
 "S1": {{true/false}},
 "S2": {{true/false}},
 "S3": {{true/false}},
 ...
}
```

---

[Begin of Webpage Contents]  
{context}  
[End of Webpage Contents]

[Begin of Statements]  
{statements}  
[End of Statements]

Figure 14: Prompt for citation-based rubric judgment in CaRR.