

Hybrid Open-Ended Tri-Evolution Makes Better Deep Researcher

Hongming Piao^{1*} Chi Liu^{1*} Mengzhuo Chen¹ Yan Shu¹ Xidong Wang¹

Derek Li¹ Ying Wei² Bryan Dai^{1†}

¹IQuest Research ²Zhejiang University
{cxiao, cliu04, cbdai}@iquestlab.com
<https://github.com/IQuestLab/ote>

Abstract

Deep research and agent evolution serve as de-facto tasks for AI agents in real-world applications toward artificial general intelligence. The former enables autonomous retrieval and integration of information in open-ended environments to tackle open-ended research tasks, yet it is constrained by the static parametric deep research capabilities of agent systems. The latter allows agents to autonomously interact with the environment to gain experiences that evolve model capabilities. However, its effectiveness has been widely validated only on verifiable tasks with standard answers, leaving a gap with open-ended research tasks. To bridge these two critical tasks, we propose the Hybrid Open-Ended Tri-Evolution (HOTE) framework, which leverages hybrid-mode reinforcement learning to facilitate the collaborative evolution of a proposer, solver and judge based on web-scale knowledge, moving toward autonomous evolving agents in open-ended tasks and environments. Extensive experiments on three long-form deep research benchmarks demonstrate that the 8B model trained via HOTE surpasses the strongest static open 8-32B models as well as those trained by state-of-the-art deep research training methods with less time overhead, and further verify that the evolution of all three modules in HOTE is indispensable.

1 Introduction

Deep research, which emphasizes autonomous handling of open-ended, long-cycle, and highly complex information retrieval and integration, has become a de-facto task for AI agents in real-world applications and a step toward artificial general intelligence (Hu et al., 2025a; OpenAI, 2025). Closed-source proprietary systems such as OpenAI Deep Research (OpenAI, 2025), Claude Research (Anthropic, 2025), Kimi-Researcher (Moonshot AI,

2025), and Grok DeepSearch (xAI, 2025) have demonstrated near-human research capabilities. Meanwhile, the open-source community has also made significant progress in building more comprehensive research workflows and end-to-end training of deep researchers capable of autonomously planning workflows (Schmidgall et al., 2025; Li et al., 2025a; Jin et al., 2025; Team et al., 2025b; Shao et al., 2025; Fang et al., 2025; Zheng et al., 2025b; Wu et al., 2025a; Yao et al., 2026; Song et al., 2025).

Although deep researchers can tackle highly complex research questions by autonomously seeking web-scale knowledge, their parameterized research capabilities are upper bounded by fixed training sets and training strategies. Autonomous interaction with the environment and evolution through experience are regarded as a path toward artificial general intelligence (Liu et al., 2025a), with self-play offering a promising paradigm for agent evolution, where an agent system learns from feedback acquired through competition with itself. For example, Huang et al. (2025); Zhao et al. (2025); Wang et al. (2025); Chen et al. (2025) proposed agent systems that act as both query proposer and solver, achieving significant results beyond the original training set or even in zero-data scenarios in domains such as mathematics, coding, or general reasoning. To address the limitation that such evolution is constrained by the agent system’s own knowledge, SPICE (Liu et al., 2025a) and Dr. Zero (Yue et al., 2026) equipped the proposer with a pretrained-scale corpus (Mahabadi et al., 2025; Yuan et al., 2025) and a search engine, respectively, taking a step forward toward evolution in open-ended environments. However, they remain limited to tasks that can be verified with deterministic answers. Given that deep researchers in real-world applications often face long-form report generation tasks without clear standard answers (Shao et al., 2025), constructing an agent evolution framework

*Equal contribution

†Corresponding author

for open-ended tasks and environments is crucial.

To fill the aforementioned gaps, we propose the Hybrid Open-ended Tri-Evolution (HOTE) framework, which consists of three co-evolving modules: proposer, solver, and judge. The solver is responsible for receiving a query, generating a research plan, conducting multi-turn information seeking, integrating information and producing a referenced research report. The judge is responsible for dynamically generating rubrics (Gunjal et al., 2025; Viswanathan et al., 2025; Shao et al., 2025) that capture the strengths and weaknesses of the solver by comparing multiple solver responses sampled for the same query, and providing rewards for the responses based on these rubrics, thereby removing the dependency on verifiable answers. The proposer is responsible for performing information seeking based on the model weaknesses identified by the judge and proposing challenging yet learnable queries. HOTE uses GRPO (Shao et al., 2024) to encourage a game between the solver and proposer, continuously improving response quality and query difficulty. Simultaneously, it employs the judge to dynamically evolve evaluation rubrics, preventing reward hacking, maintaining the learnability of difficult queries, and enabling the proposer to uncover the solver’s weaknesses. Additionally, we propose a dual-mode hybrid training strategy that includes both tool-use and no-tool modes, which achieves mutual benefit between the two modes and significantly improves training efficiency. HOTE effectively maintains the difficulty of synthetic queries during training (Figure 1(a-b)) and outperforms approaches using only the original training set within the same number of training steps (Figure 1(c)). As shown in Figure 4(a), HOTE also facilitates the collaborative progress of both the no-tool and tool-use modes.

In conclusion, our contributions are as follows:

- We propose Hybrid Open-ended Tri-Evolution (HOTE), the first deep researcher evolution framework designed for open-ended environments and open-ended tasks, bridging two paths toward artificial general intelligence: deep research and agent evolution.
- We design a co-evolution strategy for proposer, solver and judge based on reinforcement learning with hybrid modes. The strategy maintains the challenge and learnability of research tasks for the solver while avoiding reward hacking and achieving the mutual

benefit between tool-use mode and no-tool mode.

- Experimental results on three long-form deep research benchmarks demonstrate that an 8B model trained with HOTE outperforms the strongest open 8-32B models and state-of-the-art deep research training methods with less time overhead, with the co-evolution of all three modules being indispensable.

2 Method

2.1 Problem Formulation

Following (Li et al., 2025a), we build our agent on top of a concise and general ReAct framework, which provides a clear baseline for evaluating the model’s intrinsic capabilities and training strategies. The deep research model is a language model (LM) augmented with search tools. Each tool accepts a query along with its arguments and returns textual resources that can be cited in the model’s final answer. Formally, let $\mathcal{T} = \{T_1, T_2, \dots\}$ represent the set of available tools. Each tool T_k accepts a query q together with an optional argument string α , and returns an observation $o = T_k(q; \alpha)$. The model follows a policy π_θ , parameterized by θ , which generates a sequence of text s autoregressively. The sequence is initialized as $s_0 = x$, where x contains the system prompt and the task description. The model’s action space is defined as

$$\{\text{think}, \text{tool}, \text{answer}, \text{cite}\},$$

with each action associated with a corresponding protocol token. `think` (`<think>...</think>`) leverages the language model’s internal reasoning capability to plan subsequent steps based on the current state and available information. `tool` (`<call_tool name=...>...</call_tool>`) triggers the invocation of one of several search-related tools. The specific tool is selected via the name attribute, together with tool-dependent arguments omitted here. The textual output produced by the tool is appended to the context for use in later steps. `answer` (`<answer>...</answer>`) generates the final response and terminates the interaction. `cite` (`<cite id=...>...</cite>`) is embedded within the final answer to annotate claims with citation tags that reference supporting sources.

At each step i , the model samples both an action a_i and its associated content or arguments ζ_i , $(a_i, \zeta_i) \sim \pi_\theta(\cdot \mid s_i)$. If $a_i \in$

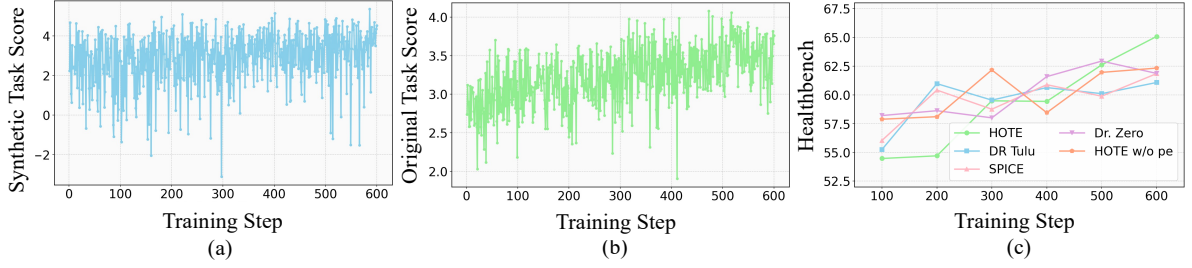


Figure 1: During the training of HOTE, (a) the scores for synthetic research tasks remain at the same level; (b) the scores for research tasks from the original training set continuously increase; (c) the scores on Healthbench surpass the baselines and maintain an upward trend.

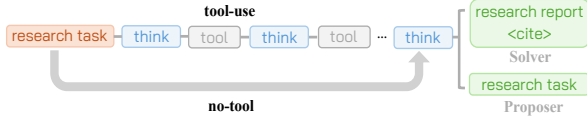


Figure 2: The inference paradigm of the solver and the proposer under tool-use and no-tool modes in HOTE.

$\{\text{think}, \text{answer}, \text{cite}\}$, the generated output ζ_i is appended to the context, yielding $s_{i+1} = s_i \oplus \langle a_i, \zeta_i \rangle$. If $a_i = \text{tool}$, the model executes the corresponding tool call, receives the observation $o_i = T_k(q_i; \alpha_i)$, where $\zeta_i = (q_i, \alpha_i)$, and updates the state as $s_{i+1} = s_i \oplus \langle a_i, \zeta_i, o_i \rangle$. This iterative procedure continues until $a_\tau = \text{answer}$, at which point ζ_τ contains the final answer. As shown in Figure 2, within HOTE, both the proposer and the solver perform inference under the same paradigm described above. **Formulating challenging research tasks for the solver constitutes a research task for the proposer itself.** The key difference is that the proposer does not include the `cite` action, since proposing research tasks does not require the presentation of citations. In HOTE, both the proposer and the solver operate under two modes: tool-use and no-tool. In the tool-use mode, the model follows the aforementioned inference paradigm. In the no-tool mode, after receiving the initial state s_0 , the model performs a single think action and then directly produces an answer action. All inference paradigms described above can be controlled through the system prompt.

2.2 Hybrid Open-ended Tri-evolution

HOTE is primarily divided into four parts: **Solver Evolution**, **Judge Evolution**, **Proposer Evolution** and **Dual-mode Hybrid Training Strategy**. Please refer to Figure 3 and Algorithm 1 for the overall framework and training pipeline.

Solver Evolution. The solver π_{θ_s} takes a re-

search task s_0 as input and, after performing think-tool interleaved reasoning, generates a long-form research report answer with `cite` represented by o . Thus, the objective of solver evolution is to make the answer better align with the research report requirements r , which also serves as the reward in the reinforcement learning and will be further discussed in the judge evolution section. We utilize GRPO (Shao et al., 2024) with token-level loss aggregation (Yu et al., 2025) to achieve solver evolution, with the goal of:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta_s) = & \mathbb{E}_{(s_0, \mathcal{R}_{s_0}) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_s^{\text{old}}}(\cdot | s_0)} \\ & \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \left(\min(r_{i,t}(\theta_s), \hat{A}_{i,t}, \right. \right. \\ & \left. \left. \text{clip}(r_{i,t}(\theta_s), \epsilon), \hat{A}_{i,t}) - \beta D_{\text{KL}}(\pi_{\theta_s} \parallel \pi_{\theta_s^{\text{ref}}}) \right) \right], \end{aligned} \quad (1)$$

$$\text{where } r_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,t} | q, o_{i,<t})}{\pi_{\theta^{\text{old}}}(o_{i,t} | q, o_{i,<t})},$$

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^G)}{\text{std}(\{r_i\}_{i=1}^G)}.$$

$\{o_i\}_{i=1}^G$ represents a group of responses to the research task s_0 . r_i denotes the reward obtained by o_i . $\mathcal{R}_{s_0}$ represents the rubric set corresponding to s_0 , which will be explained in detail in the judge evaluation section. The solver continuously progresses toward better long-form research reports based on the reward. We omit the descriptions of other symbols that can be found in Yu et al. (2025).

Judge Evolution. The judge π_{θ_j} receives a group of responses $\{o_i\}_{i=1}^G$ for s_0 from the solver as input and assigns reward r_i to each response o_i according to the rubric set $\mathcal{R}_{s_0}$ as follows:

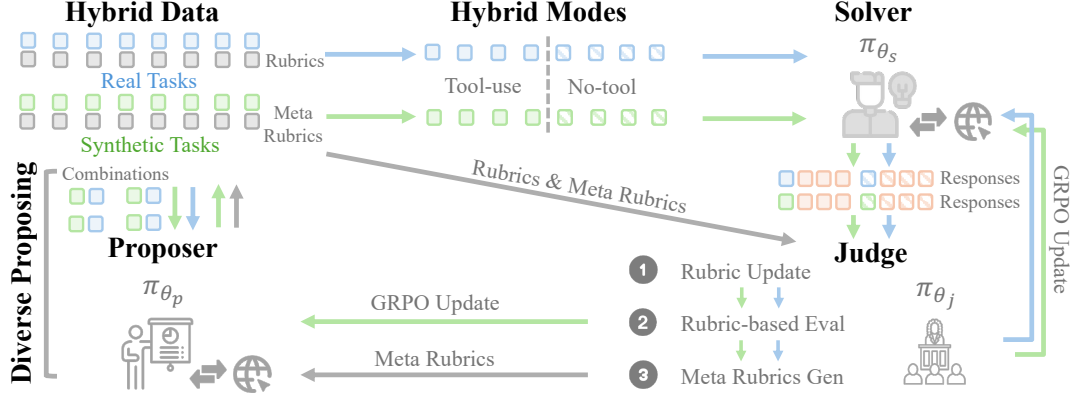


Figure 3: The overall training framework of HOTE. At each training step, we utilize hybrid data consisting of both **real tasks** and **synthetic tasks** with their corresponding persistent rubrics. Half of the tasks are configured in **tool-use** mode and the other half in **no-tool** mode. The **Solver** generates responses in hybrid mode. Based on each task’s existing rubrics and the generated responses, the **Judge** updates the rubrics, evaluates the responses and generates meta rubrics. The assessment generated by the **Judge** is used to update the **Solver**, while the portion corresponding to synthetic tasks is used to update the **Proposer**. The **Proposer** performs diverse proposing according to the meta rubrics and different combinations of tasks from the previous step, thereby generating synthetic tasks which use the meta rubrics as persistent rubrics for the next step.

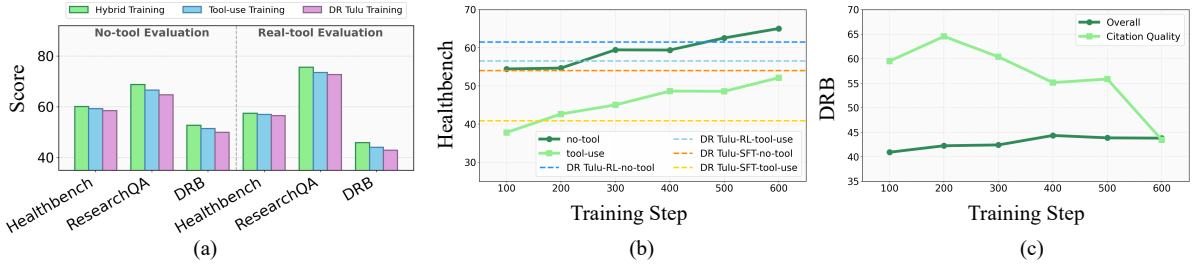


Figure 4: (a) The hybrid mode of HOTE outperforms the tool-use mode of HOTE as well as DR Tulu in both no-tool and tool-use modes; (b) Models trained with no-tool mode HOTE and DR Tulu evaluated in no-tool mode on Healthbench achieve higher scores than when evaluated in tool-use mode; (c) When trained with HOTE in no-tool mode, the scores on DRB under tool-use mode decrease after a certain number of steps.

$$r_i = \frac{\sum_{(R,w) \in \mathcal{R}_{s_0}} w \cdot \text{Judge}_{\pi_{\theta_j}}(o_i, R)}{\sum_{(R,w) \in \mathcal{R}_{s_0}} |w|}, \quad (2)$$

where R represents a rubric in $\mathcal{R}_{s_0}$ and w represents its corresponding weight. The judge’s reward for each rubric has only 0 or ± 1 . Therefore, the evolutionary objective of the judge is to provide more well-founded and discriminative rewards for the responses, ensuring the learning of the solver. As can be seen from Equation 2, the judge requires extensive inference at each step, so for training efficiency considerations, the judge in HOTE uses a fixed instruction model. In this case, the key to judge evolution shifts to how to drive the evolution of the rubric set $\mathcal{R}_{s_0}$ for s_0 . Inspired by (Shao et al., 2025), given $\mathcal{R}_{s_0} = \mathcal{R}_{s_0}^{\text{persi.}} \cup \mathcal{R}_{s_0}^{\text{active}}$ where $\mathcal{R}_{s_0}^{\text{persi.}}$ contains persistent rubrics of s_0 and $\mathcal{R}_{s_0}^{\text{active}}$

contains active rubrics of s_0 that can be deleted or added, HOTE prompts the judge to update $\mathcal{R}_{s_0}^{\text{active}}$ based on $\{o_i\}_{i=1}^G$ at each step before assigning rewards as follows:

$$\mathcal{R}_{s_0}^{\text{active}} = \text{Update}_{\pi_{\theta_j}}(s_0, \{o_i\}_{i=1}^G, \mathcal{R}_{s_0}^{\text{active}}). \quad (3)$$

The judge will generate two types of rubrics: **positive rubrics** that capture strengths or new, relevant knowledge explored by π_{θ_s} in $\{o_i\}_{i=1}^G$ but not yet reflected in $\mathcal{R}_{s_0}$, and **negative rubrics** that summarize common undesirable behaviors such as reward hacking observed across $\{o_i\}_{i=1}^G$. By observing the responses, the judge continuously tracks and uncovers weaknesses in both the rubric set and solver.

Proposer Evolution. The objective of proposer π_{θ_p} evolution is to enhance the capability to search

for materials and propose research tasks that can expose the weaknesses of the solver, based on the judge’s assessment. Similar to solver evolution, HOTE uses GRPO to achieve the evolution of the proposer as shown in Equation 1, with the distinction that s_0 becomes **proposing research tasks based on the judge’s assessment** $\mathcal{A}$ represented by s_0^p , and $\{o_i\}_{i=1}^G$ becomes **a group of research tasks proposed by the proposer** $\{o_i^p\}_{i=1}^{G'}$. However, there are two key issues left:

- The assessment from judge $\mathcal{A} = \{\text{Judge}_{\pi_{\theta_j}}(o_i, R) \mid (R, w) \in \mathcal{R}_{s_0}, 1 \leq i \leq G\}$ includes rewards for each rubric of every response, thus using all of them as input to the proposer results in excessive length, slowing down training speed.
- $\{o_i^p\}_{i=1}^{G'}$ proposed by the proposer lacks shared rubrics, making it difficult to evaluate their relative strengths and weaknesses.

Therefore, we propose **meta rubrics**, allowing the judge to summarize assessments into multiple meta rubrics, uncovering common model weaknesses among the solver’s responses as follows:

$$\mathcal{R}_{\{o_i^p\}_{i=1}^{G'}}^{\text{meta}} = \text{Meta}_{\pi_{\theta_j}}(\mathcal{A}, \mathcal{R}_{s_0}, \{o_i\}_{i=1}^G). \quad (4)$$

These meta rubrics serve as the proposer input and persistent rubrics shared across $\{o_i^p\}_{i=1}^{G'}$. On one hand, they are used for solver evolution; on the other hand, they leverage the reward of solver responses to compute the reward r_i^p of research task o_i^p as follows:

$$r_i^p = \frac{1}{M} \sum_{(R, w) \in \mathcal{R}_{\{o_i^p\}_{i=1}^{G'}}^{\text{meta}}} \mathbb{I} \cdot (1 - \mathbb{E}_{\{o_j\}_{j=1}^G \sim \pi_{\theta_s}(\cdot | o_i^p)} [\text{Judge}_{\pi_{\theta_j}}(o_j, R)]) \quad (5)$$

where $M = \left| \mathcal{R}_{\{o_i^p\}_{i=1}^{G'}}^{\text{meta}} \right|$. $\mathbb{I}$ represents whether there is an o_j that passes the rubric R . ‘1’ represents the max average reward the solver π_{θ_s} can obtain given the judge’s reward for each rubric is limited to 0 or ± 1 . Through Equation 5, we encourage the proposer to generate challenging but solvable research tasks for the solver.

2.3 Dual-mode Hybrid Training Strategy

The independent evolution of the three modules mentioned above is insufficient. They should complement one another to form an evolution pipeline

where a stronger solver stimulates a more refined judge, and a stronger proposer and judge inspire the proposer to formulate more challenging problems, which in turn train a stronger solver. As shown in Figure 3, our proposed dual-mode hybrid training strategy primarily encompasses three key features.

Hybrid Data. Except for the first step, the training data for each step (comprising a batch size of B research tasks) consists of $\frac{B}{2}$ research tasks from the original training set and $\frac{B}{2}$ synthetic research tasks proposed by the proposer based on evaluations from the previous step. Beyond leveraging existing data resources and facilitating agent evolution, this design allows synthetic research tasks generated by the proposer to be immediately solved by the solver and evaluated by the judge. The evaluation results can then be used to optimize both the proposer and the judge simultaneously, avoiding the need for repeated sampling.

Diverse Proposing. We found when the proposer generates research tasks based solely on the judge’s evaluation and all research tasks from the previous step, they tend to concentrate on the same topic, which can undermine the balance and diversity of the training data. Therefore, at each step, we prompt the proposer to generate N groups of research problems based on the judge’s evaluation and N distinct combinations of research tasks from the previous step.

Hybrid Modes. As illustrated in Figure 4(b), we found that for DR Tulu-8B-SFT, DR Tulu-8B-RL and HOTE trained solely in no-tool mode, their performance on Healthbench under no-tool mode exceeds that under tool-use mode. This phenomenon can be attributed to factors such as noise in the search tool and it is acceptable in practical applications to trade evaluation metrics for research reports with clear references. Intuitively, we think it is easier to learn research report generation techniques excluding reference searching and understanding in a no-tool training mode than a tool-use training mode. Meanwhile, as shown in Figure 4(c), for HOTE trained in no-tool mode, its performance on DRB under tool-use mode exhibits a clear pattern of initial improvement followed by decline, suggesting that no-tool training leads the model to rely excessively on parametric knowledge. Therefore, we randomly assign half of the training data in each step to no-tool mode and the other half to tool-use mode (to ensure fairness in judging synthetic research tasks, this assignment is randomized across the N groups), thereby enhancing research report

generation techniques and avoiding over-reliance on parameterized knowledge.

In actual training, we trained 600 steps using no-tool mode and then trained 700 steps using hybrid mode. Besides, we theoretically prove that the hybrid mode results in a lower expected maximum generation time in Appendix B.

3 Experiment

Our experiments aim to address five research questions. **RQ1:** Does HOTE demonstrate stronger capabilities in handling open-ended research tasks with less time overhead? **RQ2:** Are the three modules indispensable for HOTE evolution? **RQ3:** Does HOTE facilitate the collaborative progress of dual modes? **RQ4:** Is HOTE effective with different base models? **RQ5:** Does HOTE evolve more sustainably? We additionally provide the case study, the effect of judge models, prompts and diverse proposing in Appendix C and F.

3.1 Evaluations

Benchmark. We evaluated HOTE and baseline models across three long-form, open-ended benchmarks: HealthBench (Arora et al., 2025) for healthcare deep research, ResearchQA (Yifei et al., 2025) for assessing synthesis over scientific literature, the DeepResearchBench (Du et al., 2025) (DRB) for evaluating general-domain deep research tasks. For DRB, we additionally provide detailed performance across diverse aspects of the responses. DRB includes the following aspects: Comprehensiveness, Insight, Instruction Following, and Readability. In Table 1, we followed (Shao et al., 2025) by using HealthBench with 1,000 samples and ResearchQA with 776 samples. For other experimental results, we sampled 100 instances each from HealthBench and ResearchQA respectively. Please refer to Appendix G for the benchmark details.

Baselines. We compared four categories of deep researchers: *Open Deep Research Models*, including Qwen3-8B (Yang et al., 2025), Qwen3-235B-A22B (Yang et al., 2025), Search-R1-7B (Jin et al., 2025), ASearcher-Web-7B (Gao et al., 2025), WebExplorer-8B (Liu et al., 2025b), WebThinker-32B-DPO (Li et al., 2025b), Tongyi DeepResearch-30B-A3B (Team et al., 2025b); *Open Deep Research*, including DR Tulu-8B-SFT and DR Tulu-8B-RL (Shao et al., 2025); *RL Method*, including GRPO (Shao et al., 2024), GSPO (Zheng et al.,

2025a) and REINFORCE++ (Hu et al., 2025b); *Evolving Method*, including SPICE (Liu et al., 2025a) and Dr. Zero (Yue et al., 2026). We also provided *Closed Deep Research*, including Gemini 3 Pro, GPT-5, and OpenAI Deep Research; *Fixed Pipeline Deep Research*, including WebThinker QwQ-32B, WebThinker-32B-DPO, and Ai2 ScholarQA-Claude Sonnet (Singh et al., 2025) for reference. Please refer to Appendix G for implementation details.

3.2 Training Details

We utilized Qwen3-8B to initialize the checkpoint of the proposer and DR Tulu-8B-SFT to initialize the checkpoint of the solver. For Open Deep Research, RL Methods, Evolving Methods and Open Evolving Deep Research, we used the same original RL training set in DR Tulu (Shao et al., 2025) licensed under ODC-BY with 9K samples to ensure fairness. We employed Qwen3-235B-A22B-Instruct-FP8 as the judge. The batch size B was set to 48, the group size of solver G to 8, the group size of proposer G' to 6, the learning rate to $5e-7$, the maximum number of tool uses per response T to 10, the temperature to 1, and the response length to 16384. For the performance in Table 1, RL Methods and Evolving Methods were trained for 1300 steps until they converged on a held-out validation set while HOTE-8B was trained in no-tool mode for 600 steps and hybrid mode for 700 steps. We provide hyperparameter analysis in Appendix H.

3.3 RQ1: Does HOTE demonstrate stronger capabilities in handling open-ended research tasks with less time overhead?

As shown in Table 1, the HOTE-8B model surpasses the open-source solution DR Tulu on HealthBench, ResearchQA and DRB. It also outperforms Open Deep Research Models including Tongyi DeepResearch-30B-A3B. As illustrated in Figure 5, HOTE also leads existing rl and agent evolution methods. Additionally, due to the presence of hybrid mode, only half of the research tasks in the latter 700 steps out of the total 1300 steps require tool-use. Given that the maximum number of tool-use per response is T , the batch size is B and the group size is G , the maximum number of tool-use required for HOTE training is $350BTG + 175B$ (term 1 for solver, term 2 for proposer). In contrast, for DR Tulu it is $1900BTG$ while for RL Methods and Evolving Methods it is $1300BTG$. Furthermore, as indicated in Table 2, even with the addition of

¹We use the 1900-step checkpoint of DR Tulu.

Table 1: Performance comparison across long-form deep research benchmarks. HOTE-8B outperforms existing *Open Deep Research Models*, *Open Deep Research*, *RL Methods* and *Evolving Methods*.

Method	HealthBench	ResearchQA	DRB					Average
			Overall	Comp	Insight	Instruction	Readability	
Closed Deep Research								
Gemini 3 Pro + Search	38.0	74.3	46.3	43.4	44.9	49.8	49.0	52.9
GPT-5 + Search	59.5	78.2	50.7	26.7	21.3	41.0	29.4	62.8
OpenAI Deep Research	53.8	79.2	46.9	46.8	45.2	49.2	47.1	60.0
Open Deep Research Models								
Qwen3-8B	5.9	46.3	18.2	14.3	8.7	29.5	24.4	23.5
Qwen3-235B-A22B	21.3	50.7	22.5	19.1	17.3	30.6	25.1	31.5
Search-R1-7B	-0.1	27.9	9.5	5.2	2.1	18.6	16.8	12.4
ASearcher-Web-7B	-13.0	19.4	7.8	5.1	1.7	15.2	11.8	4.7
WebExplorer-8B	33.7	64.8	36.7	33.7	28.5	45.7	42.2	45.1
WebThinker-32B-DPO	11.1	48.6	23.3	19.7	12.3	36.8	26.3	27.7
Tongyi DeepResearch-30B-A3B	46.2	66.7	40.6	39.1	34.3	46.8	45.4	51.2
Fixed Pipeline Deep Research								
WebThinker QwQ-32B (report)	36.5	72.8	37.9	36.2	32.6	43.2	42.9	49.1
WebThinker-32B-DPO (report)	39.4	74.2	40.6	39.4	35.4	46.0	43.5	51.4
Ai2 ScholarQA-Claude Sonnet (report)	32.0	75.0	36.1	35.1	32.0	40.5	38.9	47.7
Open Deep Research								
DR Tulu-8B-SFT	38.1	68.5	39.0	36.3	35.3	45.5	39.5	48.5
DR Tulu-8B-RL ¹	50.2	74.3	43.4	41.7	41.8	48.2	41.3	56.0
RL Methods								
GRPO	49.6	73.5	43.1	40.8	42.1	46.9	42.6	55.4
GSPO	51.0	75.1	43.6	42.5	40.9	47.3	43.7	56.6
REINFORCE++	50.8	74.8	43.1	41.2	42.7	46.1	42.4	56.2
Evolving Methods								
SPICE-8B	50.2	73.9	42.1	40.6	40.9	46.1	40.8	55.4
Dr. Zero-8B	52.1	73.2	43.7	41.5	42.1	46.5	44.7	56.3
Open Evolving Deep Research								
HOTE-8B	54.4	76.9	45.9	44.9	45.4	47.8	45.8	59.1

Table 2: Average training time per step for baselines, no-tool mode and hybrid mode of HOTE.

Method	Wall-clock (s/step)	GPU (h/step)
DR Tulu	1136.8	3.8
GRPO	1024.2	3.5
GSPO	1011.3	3.4
REINFORCE++	979.6	3.3
SPICE	1313.7	4.0
Dr. Zero	1301.0	4.0
HOTE no-tool	382.0	1.5
HOTE hybrid	753.3	2.6

proposer evolution, both the no-tool mode in the first 600 steps and the hybrid mode in the latter 700 steps contribute to improvements in training speed.

3.4 RQ2: Are the three modules indispensable for HOTE evolution?

We compared HOTE, SPICE, the HOTE version without judge evolution (HOTE w/o je, equivalent to Dr. Zero using rubric-based reward and GRPO), and the HOTE version without proposer evolution (HOTE w/o pe, the proposer’s parameters are fixed) in the no-tool mode. As shown in Figure 5, when training in the no-tool mode using HOTE, although HOTE initially performed slightly worse on the benchmark compared to SPICE, HOTE

w/o je and HOTE w/o pe, it gradually achieved overall superiority as training progressed. More importantly, while HOTE w/o je, HOTE w/o pe and SPICE approached convergence, HOTE maintained a stronger upward trend. Moreover, as can be seen from Figure 6, with proposer evolution enabled, the scores of synthetic research tasks are more stable compared to fixed proposer parameters, which indicates that proposer evolution helps maintain the difficulty of research tasks.

3.5 RQ3: Does HOTE facilitate the collaborative progress of dual modes?

Figure 4(a) shows the performance of HOTE, the open-source training approach DR Tulu, as well as HOTE trained exclusively in tool-use mode, evaluated on HealthBench, ResearchQA, and DRB under both no-tool and tool-use modes. HOTE outperforms both DR Tulu and the single-mode version across both no-tool and tool-use evaluation modes, achieving collaborative progress in the dual modes by enhancing research report generation techniques while avoiding over-reliance on parameterized knowledge.

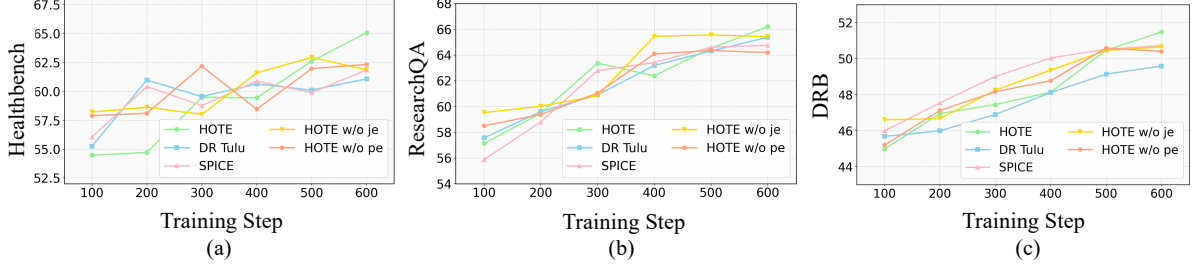


Figure 5: In HealthBench (a), ResearchQA (b), and DeepResearchBench (c), after 600 steps of training in no-tool mode, HOTE outperforms SPICE, HOTE w/o je and HOTE w/o pe while demonstrating an upward trend.

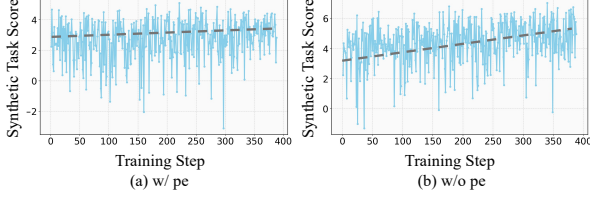


Figure 6: (a) During the training in no-tool mode with proposer evolution enabled, the solver’s synthetic task score remains stable, indicating that proposer evolution maintains the challenge of the tasks for the evolving solver; (b) After disabling proposer evolution, the solver’s synthetic task score gradually increases.

Table 3: Performance comparison with Llama3.1-8B-Instruct supervised fine-tuned by dr-tulu-sft-data as the base model.

Method	HealthBench	ResearchQA	DRB
DR Tulu	28.8	60.0	32.3
GRPO	27.2	59.8	31.0
GSPO	27.8	59.0	31.5
REINFORCE++	27.2	59.3	32.6
SPICE	29.2	62.0	32.9
Dr. Zero	28.4	61.0	31.5
HOTE	33.1	64.2	35.1

3.6 RQ4: Is HOTE effective with different base models?

As shown in Table 3, we also provided the performance comparison on Llama3.1-8B-Instruct supervised fine-tuned by dr-tulu-sft-data (Shao et al., 2025) for 5 epochs. HOTE maintains its lead across three benchmarks and over baselines that we can train by ourselves including *Open Deep Research*, *RL Methods* and *Evolving Methods*. The absolute scores are lower than using DR Tulu-8B-SFT fine-tuned from Qwen3-8B due to the lower capability of the base model.

3.7 RQ5: Does HOTE evolve more sustainably than baselines?

We compared the performance of HOTE and the baselines during training from 1200 to 1500 to-

Table 4: Average performance comparison across three benchmarks between HOTE and baselines from 1200 steps to 1500 steps in total. HOTE evolves more sustainably than baselines.

Method	1200	1300	1400	1500
GRPO	55.4	55.4	55.2	55.2
GSPO	56.4	56.6	56.6	56.4
REINFORCE++	56.2	56.2	55.7	56.0
SPICE	55.5	55.4	55.2	55.4
Dr. Zero	56.3	56.3	56.5	56.1
HOTE	58.0	59.1	59.6	59.9

tal steps. We use the average performance on the three benchmarks. As shown in Table 4, the baselines have already converged, whereas HOTE not only outperforms the baselines but also continues to exhibit an upward trend. HOTE can sustain continuous evolution for at least 252 hours (1500 steps) of wall-clock time.

4 Conclusion

We propose Hybrid Open-Ended Tri-Evolution (HOTE), aiming to develop a deep researcher capable of autonomous evolution in open-ended environments for open-ended tasks with less time overhead. Through a well-designed reinforcement learning with hybrid modes, HOTE achieves synergistic evolution among the proposer, solver and judge as well as the mutual benefit between no-tool and tool-use modes. On three long-form deep research benchmarks, HOTE-8B outperforms the strongest open 8-32B models and state-of-the-art deep research training methods with less time overhead. In future work, we will continue to explore how to handle noise in real-world search tools during the evolutionary process, how to break free from dependence on original training dataset and how to scale HOTE to larger MoE models.

Limitations

The evolution gradually slows down as training progresses and is difficult to obtain perfect scores, suggesting that the upper bound of evolution may still be constrained by model scale. Investigating the scaling capability of HOTE will be a major direction of our future work. The proposed method still relies on the initial training data, but we believe that transcending the limitations of existing training data through evolution is inherently valuable.

References

- Anthropic. 2025. Claude takes research to new places. <https://www.anthropic.com/news/research>. Accessed: 2025-04.
- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.
- Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Self-questioning language models. *arXiv preprint arXiv:2508.03682*.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.
- Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.
- FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.
- Tianqing Fang, Zhisong Zhang, Xiaoyang Wang, Rui Wang, Can Qin, Yuxuan Wan, Jun-Yu Ma, Ce Zhang, Jiaqi Chen, Xiyun Li, and 1 others. 2025. Cognitive kernel-pro: A framework for deep research agents and agent foundation models training. *arXiv preprint arXiv:2508.00414*.
- 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. *arXiv preprint arXiv:2508.07976*.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.
- Chen Hu, Haikuo Du, Heng Wang, Lin Lin, Mingrui Chen, Peng Liu, Ruihang Miao, Tianchi Yue, Wang You, Wei Ji, and 1 others. 2025a. Step-deepresearch technical report. *arXiv preprint arXiv:2512.20491*.
- Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025b. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*.
- Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. 2025. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-rl: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.
- Kuan Li, Zhongwang Zhang, Huifeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litu Ou, Dingchu Zhang, Xixi Wu, Jialong Wu, and 1 others. 2025a. Websailor-v2: Bridging the chasm to proprietary agents via synthetic data and scalable reinforcement learning. *arXiv preprint arXiv:2509.13305*.
- Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025b. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.
- Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. 2025a. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*.
- Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, and 1 others. 2025b. Webexplorer: Explore and evolve for training long-horizon web agents. *arXiv preprint arXiv:2509.06501*.
- Rabeeh Karimi Mahabadi, Sanjeev Satheesh, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-cc-math: A 133 billion-token-scale high quality math pretraining dataset. *arXiv preprint arXiv:2508.15096*.
- Moonshot AI. 2025. Kimi-researcher: End-to-end rl training for emerging agentic capabilities. <https://moonshotai.github.io/Kimi-Researcher/>.

- OpenAI. 2025. Introducing deep research. <https://openai.com/index/introducing-deep-research/>. Accessed: 2025-02.
- Tianrui Qin, Qianben Chen, Sinuo Wang, He Xing, King Zhu, He Zhu, Dingfeng Shi, Xinxin Liu, Ge Zhang, Jiaheng Liu, and 1 others. 2025. Flash-searcher: Fast and effective web agents via dag-based parallel execution. *arXiv preprint arXiv:2509.25301*.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.
- 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. 2025. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, and 1 others. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.
- Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D Hwang, Jason Dunkleberger, and 1 others. 2025. Ai2 scholar qa: Organized literature synthesis with attribution. *arXiv preprint arXiv:2504.10861*.
- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.
- MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, and 1 others. 2025a. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*.
- Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025b. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.
- Gerald Tesauro and 1 others. 1995. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68.
- Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. 2025. Checklists are better than reward models for aligning language models. *arXiv preprint arXiv:2507.18624*.
- Yuxuan Wan, Tianqing Fang, Zaitang Li, Yintong Huo, Wenxuan Wang, Haitao Mi, Dong Yu, and Michael R Lyu. 2026. Inference-time scaling of verification: Self-evolving deep research agents via test-time rubric-guided verification. *arXiv preprint arXiv:2601.15808*.
- Yinjie Wang, Ling Yang, Ye Tian, Ke Shen, and Mengdi Wang. 2025. Cure: Co-evolving coders and unit testers via reinforcement learning. In *The Thirtieth Annual Conference on Neural Information Processing Systems*.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.
- Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, and 1 others. 2025a. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*.
- Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and 1 others. 2025b. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*.
- Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025c. Agentic reasoning: A streamlined framework for enhancing llm reasoning with agentic tools. *arXiv preprint arXiv:2502.04644*.
- Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.
- xAI. 2025. Grok 3 beta — the age of reasoning agents. <https://x.ai/news/grok-3>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

- Yi Yao, He Zhu, Piaohong Wang, Jincheng Ren, Xinlong Yang, Qianben Chen, Xiaowan Li, Dingfeng Shi, Jiaxian Li, Qiexiang Wang, and 1 others. 2026. O-researcher: An open ended deep research model via multi-agent distillation and agentic rl. *arXiv preprint arXiv:2601.03743*.
- Li S Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. 2025. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *arXiv preprint arXiv:2509.00496*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.
- Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Ilya Kulikov, Kyunghyun Cho, Dong Wang, Yuandong Tian, Jason E Weston, and 1 others. 2025. Naturalreasoning: Reasoning in the wild with 2.8 m challenging questions. *arXiv preprint arXiv:2502.13124*.
- Zhenrui Yue, Kartikeya Upasani, Xianjun Yang, Suyu Ge, Shaoliang Nie, Yuning Mao, Zhe Liu, and Dong Wang. 2026. Dr. zero: Self-evolving search agents without training data. *arXiv preprint arXiv:2601.07055*.
- Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025a. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.
- Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025b. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.

A Related Work

We summarize the contribution of HOTE in Table 5.

A.1 Deep Research Agents

Deep research, defined as AI agents’ capability to handle open-ended, long-term, and highly complex information retrieval and integration, has become key for AI agents to move beyond conversational interaction toward general autonomy (Hu et al., 2025a; OpenAI, 2025).

On the inference front, Wu et al. (2025c); Qin et al. (2025); Schmidgall et al. (2025) have shown that constructing complex workflows and context management can lead to substantial performance improvements. However, such methods rely on manual prompting, lack generality and flexibility, and make it difficult to evaluate the inherent autonomous agent capabilities of the model (Li et al., 2025a). On the training front, research has primarily focused on how to end-to-end train autonomous deep research agents based on flexible reasoning paradigms similar to ReAct (Yao et al., 2022), enabling them to self-plan, acquire knowledge and summarize. Search-R1 (Jin et al., 2025) applies reinforcement learning with verifiable rewards (RLVR) to enhance search capabilities and is trained mainly on short-form question answering (Wei et al., 2024; Wu et al., 2025b; Ho et al., 2020). This approach has been explored in many recent follow-up studies, including WebExplorer (Liu et al., 2025b), Tongyi Deep Research (Team et al., 2025b) and WebSailor-V2 (Li et al., 2025a). WebThinker (Li et al., 2025b) and MiroThinker (Team et al., 2025a) extend training to longer report generation and more rounds of tool usage. To address the lack of clearly defined evaluation metrics for long-form deep research responses, DR Tulu (Shao et al., 2025) proposes Reinforcement Learning via Evolving Rubrics (RLER), which dynamically updates evaluation rubrics based on sampled policy responses. Although the above studies enable agents to autonomously conduct research based on user queries, they lack a process for autonomous exploration and improvement of deep research capabilities. Dr. Zero (Yue et al., 2026) designs a framework based on search-based proposer–solver self-play, enabling the two to co-evolve without exposure to any training data, but it is limited to short-form and easily verifiable question answering.

Therefore, we propose the first deep research agent evolution framework that supports open-ended long-form report generation tasks, aiming to achieve both practicality and autonomy simultaneously.

A.2 Agent Evolving with Self-play

Agent evolution has long been regarded as a pathway toward achieving artificial general intelligence, signifying the capability of agents to autonomously interact with the environment and continuously learn (Liu et al., 2025a). Self-play offers a highly promising paradigm for agent evolution, wherein an agent system learns from feedback automatically generated through competition with itself. In the domain of games, self-play has led to achievements such as TD-Gammon’s backgammon mastery (Tesauro et al., 1995), AlphaGo’s superhuman performance in Go (Silver et al., 2017), and CICO’s capability to understand cooperative strategies (FAIR et al., 2022). In the field of large language models, some approaches enable models to serve dual roles as solver and judge, optimizing strategies without the need for human annotation (Chen et al., 2024; Wu et al., 2024; Yuan et al., 2024; Wan et al., 2026). However, such evolution is constrained by the queries in the training set, limiting the model’s ability to autonomously explore new knowledge and skills. By assigning the agent system the roles of both query proposer and solver, significant improvements have been achieved in areas such as mathematics, coding, and general reasoning, surpassing the limitations of the original training set and even demonstrating zero-data effectiveness (Huang et al., 2025; Zhao et al., 2025; Wang et al., 2025; Chen et al., 2025). To further overcome the inherent limitations of agent capabilities, methods such as SPICE (Liu et al., 2025a) and Dr. Zero (Yue et al., 2026) provide proposers with large-scale corpora and search engines, facilitating the evolution of agent systems in open-ended environments. However, existing approaches remain confined to verifiable tasks, falling short of addressing the reality of numerous open-ended tasks with ambiguous or undefined boundaries encountered in real-world applications.

Therefore, we propose an open-ended agent evolution framework tailored to open-ended tasks that are difficult to verify. Through mutual play among proposer, solver and judge, the framework enables collaborative evolution with web-scale knowledge.

Table 5: The contribution of HOTE.

Method	Proposer Evolve	Solver Evolve	Judge Evolve	Open-ended Task	Open-ended Environment
Dr. zero (Yue et al., 2026) (Meta)	✓	✓	✗	✗	✓
DR Tulu (Shao et al., 2025) (Ai2)	✗	✓	✓	✓	✗
Spice (Liu et al., 2025a) (ICLR 2026)	✓	✓	✗	✗	✗
R-zero (Huang et al., 2025) (ICLR 2026)	✓	✓	✗	✗	✗
HOTE	✓	✓	✓	✓	✓

B Proof of Expected Maximum Generation Time Comparison

We formally derive the inequality between the expected maximum generation time of a **tool-use** strategy and a **hybrid-mode** strategy.

B.1 Problem Setup

Let X denote the random variable representing the generation time for the **tool-use** mode, and Y denote the generation time for the **no-tool** mode. We assume these follow normal distributions with identical variances σ^2 but distinct means:

$$X \sim \mathcal{N}(\mu_T, \sigma^2), \quad (6)$$

$$Y \sim \mathcal{N}(\mu_N, \sigma^2), \quad (7)$$

where $\mu_T > \mu_N$. Let $F_X(t)$ and $F_Y(t)$ denote the cumulative distribution functions (CDFs) of X and Y , respectively. Since $\mu_T > \mu_N$ and the variances are equal, we have the strict inequality for the CDFs:

$$F_X(t) < F_Y(t), \quad \forall t \in \mathbb{R}. \quad (8)$$

We consider the number of generations of the solver as K .

- **Strategy A (tool-use):** The maximum generation time M_A is defined as the maximum of K independent and identically distributed (i.i.d.) variables $X_1, \dots, X_K \sim X$:

$$M_A = \max\{X_1, \dots, X_K\}. \quad (9)$$

- **Strategy B (hybrid mode):** The maximum generation time M_B is defined as the maximum of $K/2$ variables of type X and $K/2$ variables of type Y , all mutually independent:

$$M_B = \max\{X_1, \dots, X_{K/2}, Y_1, \dots, Y_{K/2}\}. \quad (10)$$

Theorem B.1. *The expected maximum generation time of Strategy A is strictly greater than that of Strategy B, i.e., $E[M_A] > E[M_B]$.*

Proof. First, we derive the cumulative distribution functions for the random variables M_A and M_B . For any $t \in \mathbb{R}$, the probability that the maximum of a set of independent variables is less than or equal to t is the product of their individual probabilities.

For Strategy A:

$$P(M_A \leq t) = \prod_{i=1}^K P(X_i \leq t) = [F_X(t)]^K. \quad (11)$$

For Strategy B:

$$P(M_B \leq t) = \left(\prod_{i=1}^{K/2} P(X_i \leq t) \right) \cdot \left(\prod_{j=1}^{K/2} P(Y_j \leq t) \right) = [F_X(t)]^{K/2} [F_Y(t)]^{K/2}. \quad (12)$$

Using the inequality from Eq. (8), where $F_X(t) < F_Y(t)$ for all t , and noting that $F_X(t) > 0$ for sufficiently large t , we compare the two probabilities:

$$\begin{aligned} P(M_B \leq t) &= [F_X(t)]^{K/2} [F_Y(t)]^{K/2} \\ &> [F_X(t)]^{K/2} [F_X(t)]^{K/2} \\ &= [F_X(t)]^K \\ &= P(M_A \leq t). \end{aligned} \quad (13)$$

Thus, $P(M_B \leq t) > P(M_A \leq t)$ for all t where $F_X(t) > 0$. This implies that M_A stochastically dominates M_B (first-order stochastic dominance).

In terms of the survival function (tail probability), this inequality is reversed:

$$P(M_A > t) = 1 - P(M_A \leq t) \quad (14)$$

$$> 1 - P(M_B \leq t) \quad (15)$$

$$= P(M_B > t). \quad (16)$$

The expected value of a random variable Z can be expressed as the integral of its survival function

over its support. Assuming the support covers the real line:

$$E[Z] = \int_{-\infty}^{\infty} t f_Z(t) dt \quad (17)$$

$$= \int_0^{\infty} P(Z > t) dt - \int_{-\infty}^0 P(Z \leq t) dt. \quad (18)$$

Given the stochastic dominance established above, the strict inequality holds for the expectation:

$$E[M_A] > E[M_B]. \quad (19)$$

□

C Case Study of Generated Reports

We conducted case studies on HOTE-8B and DR Tulu-8B to illustrate the advantages of HOTE. We omit the think and tool because of they are too long. In practical applications, the final research report will be additionally appended with the searched references. In Case 1, as shown in Figure 7-8, HOTE demonstrates: (a) more comprehensive information: the response from HOTE-8B provides detailed citations from EACS guidelines, including baseline examination items (viral load, CD4 count, complete blood count, metabolic indicators, TB screening, opportunistic infection assessment, etc.); (b) better structure: the answer is clearly organized into sections such as "Summary", "Baseline workup" and "Virologic check points"; (c) stronger contextual awareness: it correctly identifies that this is a question for medical professionals, offering detailed guidelines suitable for their level. In contrast, DR Tulu offers a more concise response, presenting only "Bottom line" recommendations and lacking a complete monitoring timeline and baseline examination details. In Case 2, as shown in Figure 10-14, HOTE (a) correctly identifies an emergency: clearly states that "acute angle-closure glaucoma is a true ophthalmic emergency" requiring "immediate evaluation and treatment to prevent rapid, irreversible vision loss"; (b) provides specific action advice: explains what the patient should do (seek evaluation by an ophthalmologist) and what examinations the doctor will perform; (c) offers complete clinical information: including symptom descriptions (severe eye pain, blurred vision, halos, headache, nausea) and treatment methods (laser peripheral iridotomy). DR Tulu, while providing background

medical knowledge, fails to clearly inform the patient that this is an emergency requiring immediate medical attention.

D Case Study of Synthetic Tasks

We added a case-based analysis of whether the synthetic tasks are well-posed, diverse, genuinely challenging, and aligned with the meta-rubrics. **1. CaCl₂ task:** Calculate the molarity from 15.0 g of CaCl₂ and 200 mL of solution, and cite the atomic-weight sources. **2. EV-battery task:** Given a 100 kWh capacity, a cell cost of 650 RMB/kWh, and 200 kg of lithium carbonate priced at 100,000 RMB/ton, calculate the lithium-material cost and its share of the battery cost, and discuss price trends through 2026. **3. TTP task:** A 38-year-old woman has fatigue, dysuria, melena, thrombocytopenia, and microangiopathic hemolytic anemia but no acute kidney injury; determine the most likely diagnosis and the next diagnostic and management steps.

The tasks are **well-posed** through explicit inputs and evaluation requirements. Task 1 has a uniquely verifiable answer of 0.676 M, while the lithium-material cost and cost share in Task 2 are 20,000 RMB and approximately 30.8%, respectively. For the open-ended Task 3, the prompt explicitly requires diagnosis, testing, and management. Its completeness and correctness can therefore be evaluated against predefined answer components.

The tasks are **diverse in both domain and reasoning format**. They span chemistry, energy economics, and hematology while eliciting numerical calculation, cost modeling, and diagnostic reasoning rather than superficial variants of one task type.

Their difficulty **does not arise merely from increased length or obscure knowledge**. Task 1 requires the final result to remain consistent with every intermediate calculation. Task 3 uses absent acute kidney injury as a distractor: the model must recognize that it does not exclude TTP and distinguish ADAMTS13-based confirmation from the need for immediate plasma exchange.

The tasks are **aligned with meta rubrics**. Task 1 → Multi-Step Numerical Consistency; Source-Grounded Factual Precision: It requires consistency across unit conversion, molar mass, moles, and final concentration, together with reliable sources for the atomic weights. Task 2 → Multi-Constraint Quantitative Synthesis; Temporal Evidence Grounding: It requires integrating capac-

ity, unit cost, material mass, and material price using traceable and temporally relevant evidence. Task 3 → Robust Reasoning under Misleading Evidence; Complete Multi-Part Decision Making: It tests whether a missing canonical feature is incorrectly treated as exclusionary and requires diagnosis, confirmatory testing, and urgent management.

E Algorithm

We provide the complete training process of HOTE in Algorithm 1.

F The Effect of Judge Models for Training, Prompts and Proposing

We further trained with Qwen3-235BA22B-Think and Qwen3-30BA3B-Instruct as the judge model (2507-FP8 version), along with the average wall-clock time per step. The results in Table 6 show that: (i) a smaller-scale judge model leads to a moderate performance degradation; (ii) a thinking model achieves nearly identical performance but substantially reduces training efficiency. Therefore, we recommend using large-scale open-source instruct models to strike a balance between effectiveness and computational overhead. We consistently set: temperature=0, max_tokens=16384, top_p=1.0.

The prompts are role-defining system instructions for the proposer, solver, and judge modules, designed to specify each module’s task and output format. The samples are minimal format demonstrations and are not part of the evaluation benchmarks. To test sensitivity, we replaced samples and rephrased role-defining instructions with three different sets on HOTE and three baselines. We observed negligible impact on final performance in Table 9, suggesting that the method is not materially dependent on a particular prompt/sample choice.

As shown in Table 7, the diverse proposing effectively improve the performance on three benchmarks, illustrating its importance in ensuring the quality of proposed research tasks.

G Details

G.1 Implementation details

For *RL Methods* and *Evolving Methods* that we can fully control the training process, since long-form deep research tasks do not have standard reference answers, we consistently adapted them from RLVR to rubric-based reward following (Shao

et al., 2025) without judge evolution. For *Evolving Methods* including SPICE and Dr. Zero, we also consistently adapted them in the same manner as HOTE by utilizing Qwen3-8B to initialize the proposer checkpoint and DR Tulu-8B-SFT to initialize the solver checkpoint. For *Open Deep Research Models*, *Open Deep Research*, *RL Methods*, *Evolving Methods* and HOTE that we can fully control the inference process, we use Serper API for google_search, Jina API for web_browse and Semantic Scholar API for paper_search. We ensured that no data from the benchmark was added to the training set, and we also blocked search tools from accessing the benchmark website. For *Closed Deep Research* that we cannot fully control the training and inference process, we also provide their results for reference following (Shao et al., 2025) but not for a strict comparison with them. We plan to fully release our models and codes upon acceptance.

G.2 Benchmark details

Judge. To avoid the model simply using the biases of the judge during training, and also to follow the official evaluation of HealthBench (Arora et al., 2025), DRB (Du et al., 2025) and ResearchQA (Yifei et al., 2025), different judge models were employed for different benchmarks: GPT-4.1 was used for Healthbench; Gemini-2.5-flash for DRB; and GPT-4.1-mini for ResearchQA. Higher scores consistently indicate better quality across all benchmarks. HealthBench calculates a normalized score based on physician-created rubrics that reward desired behaviors and penalize undesirable ones; ResearchQA measures the thoroughness of addressing literature-derived criteria on a 0-100% scale; and DRB computes a macro-average score across four quality dimensions via comparison against high-quality reference reports. Please refer to the references for specific implementation.

Reliability. We run the evaluation of HOTE on all three benchmarks three times. As shown in Table 8, the standard deviations are small and HOTE consistently maintains its lead. LLM-as-a-judge can provide a stable evaluation for the three benchmarks. Besides, human experts are substantially involved in rubric design across all three benchmarks to ensure the reliability: HealthBench uses conversation-specific rubrics written by 262 physicians, with consensus criteria added only when a majority of reviewing physicians agree they are relevant; ResearchQA derives query-specific rubrics from expert-written survey sections and further

Table 6: Performance and training efficiency under different judge models.

Judge Model	HealthBench	ResearchQA	DRB	Avg. Wall-clock Second/Step
Qwen3-235BA22B-Think	54.6	76.9	45.5	1065.5
Qwen3-30BA3B-Instruct	50.2	72.2	42.0	723.7
Qwen3-235BA22B-Instruct	54.4	76.9	45.9	753.3

Table 7: The effect of diverse proposing.

Benchmark	w/ diverse proposing	w/o diverse proposing
HealthBench	54.4	50.2
ResearchQA	76.9	74.1
DRB	45.9	42.2

validates them with 31 Ph.D. annotators across 8 fields; and DRB builds on tasks crafted and iteratively refined by domain experts, while its adaptive criteria are anchored in four top-level dimensions established from domain expertise: comprehensiveness, insight, instruction-following, and readability.

The choice of benchmark. The chosen benchmarks are for a complementary evaluation in different domains: they evaluate distinct aspects of long-form deep research quality including (Healthbench) health-related safety and communication quality, (Researchqa) scholarly synthesis across 7 research domains (Life & Earth Sciences, Engineering & Computer Science, Physical Sciences, Health Sciences & Medicine, Social Sciences, Humanities, Economics), and (DRB) end-to-end deep-research report quality across 22 domains (Science & Technology, Finance & Business, Software Development, Education & Jobs, Health, Literature, History, Hardware, Industrial, Art & Design, Games, Crime & Law, Entertainment, Sports & Fitness, Software, Transportation, Religion, Home & Hobbies, Travel, Food & Dining, Fashion & Beauty, Social Life) rather than a single narrow criterion.

H Hyperparameter Analysis

We used HealthBench, ResearchQA and DRB to analyze the impact of different batch sizes B , solver group sizes G , proposer group sizes G' , and the numbers of training steps in no-tool mode and tool-use mode. As shown in Table 10, increasing B , G , and G' first improves performance and then leads to a plateau. Therefore, we select $B = 48$, $G = 8$, and $G' = 6$. We further conduct the tool-use training until convergence after different steps of no-tool training to explore the effect of no-tool steps. As shown in Table 10, increasing the number of no-tool training steps first improves performance

and then causes a decline, possibly because the model becomes overly reliant on parametric knowledge as training progresses. Therefore, we train the no-tool mode for 600 steps. For the learning rate, maximum number of tool uses per response, temperature, and response length, we reused the hyperparameter settings ablated in (Shao et al., 2025).

I Specific Prompts

Figure 15 and Figure 16 show the system prompts for the solver in tool-use mode. Figure 17 shows the system prompt for the solver in no-tool mode. Figure 18 and Figure 19 show the system prompts for the proposer in tool-use mode. Figure 20 shows the user prompt for the proposer in tool-use mode. Figure 21 and Figure 22 show the system prompt and user prompt for the proposer in no-tool mode, respectively. Figure 23 and Figure 24 show the system prompts for the judge updating rubrics according to Equation 3. Figure 25 shows the system prompt for the judge assigning rewards based on rubrics. Figure 26 and Figure 27 show the system prompts for the judge generating meta rubrics.

J Standalone Evaluation of Judge

We added a standalone blinded evaluation spanning early (200 steps), middle (400 steps), and late training (600 steps) during real-tool training. In a preliminary evaluation of 18 high vs. low-scored response pairs (6 pairs randomly sampled for each step), the majority preference of three experts agreed with the evolving judge on 16/18 pairs.

For example, the rubrics below were newly generated during training rather than taken from the original static rubrics. The reward of those evolving rubrics effectively differentiated answers of varying quality.

Table 8: Performance statistics across three evaluation runs.

Benchmark	Mean $\pm$ Std
HealthBench	54.5 $\pm$ 0.1
ResearchQA	76.7 $\pm$ 0.2
DRB	45.9 $\pm$ 0.0

Table 9: Performance comparison across different samples and role-defining instructions.

Exp	Method	HealthBench	ResearchQA	DRB
Different samples	Search-R1-7B	0.0 $\pm$ 0.0	28.0 $\pm$ 0.2	9.5 $\pm$ 0.0
	WebExplorer-8B	33.7 $\pm$ 0.1	64.8 $\pm$ 0.2	36.7 $\pm$ 0.0
	Tongyi DeepResearch-30B-A3B	46.2 $\pm$ 0.4	66.7 $\pm$ 0.1	40.6 $\pm$ 0.1
	HOTE-8B	54.4$\pm$0.2	76.6$\pm$0.2	46.3$\pm$0.3
Different role-defining instructions	Search-R1-7B	0.0 $\pm$ 0.0	27.9 $\pm$ 0.2	9.5 $\pm$ 0.0
	WebExplorer-8B	33.8 $\pm$ 0.1	64.8 $\pm$ 0.1	36.7 $\pm$ 0.0
	Tongyi DeepResearch-30B-A3B	46.2 $\pm$ 0.1	66.6 $\pm$ 0.2	40.6 $\pm$ 0.0
	HOTE-8B	54.4$\pm$0.1	76.7$\pm$0.2	45.9$\pm$0.1

- CaCl₂ molarity. The evolved negative rubric, “Numerical Result–Process Inconsistency,” penalizes a final answer that contradicts its intermediate calculation. Two additional rubrics, “Hydration State Awareness and Justification” and “Uncertainty Propagation from Atomic Weights,” assess whether the response states its anhydrous/hydrated-salt assumption and handles atomic-weight precision appropriately. One response correctly calculated (0.676) M but incorrectly reported (0.772) M as its final result and received 1.0, whereas a response consistently reporting (0.676) M received 3.5. All three experts agreed with this ranking.
- Natalizumab-associated PML. The evolved rubric “Pathognomonic Imaging Sign Integration” requires the response to connect PML-specific imaging findings to the clinical presentation rather than merely mentioning generic white-matter lesions. “Temporal-Immunological Imaging Interpretation” evaluates whether the response distinguishes classic PML from enhancement and edema caused by PML–IRIS after natalizumab discontinuation. The negative rubric “Misattribution of PML-Specific Imaging to MS” penalizes interpreting characteristic PML findings as ordinary multiple-sclerosis plaques. A complete response connecting natalizumab exposure, JC virus, asymmetric neurological deficits, and MRI findings received 8.8, whereas an incomplete response without a definitive diagnosis received 0.4. All experts agreed with this ordering.
- Post-treatment reaction in lepromatous leprosy. “Discriminative Mechanistic Integration” requires the response to jointly consider onset timing, immunological mechanism, and the distinction between Type 1 and Type 2 leprosy reactions. “Pathophysiological Category Discrimination” distinguishes inflammation caused by antigen release from classical adaptive hypersensitivity, while the negative rubric “Conflation of Baseline Pathology with Treatment Reaction” penalizes equating the chronic immune profile of lepromatous leprosy with the acute post-treatment reaction. A response incorrectly identifying a Type 1 leprosy reaction and Type IV hypersensitivity received 1.0. A response correctly identifying erythema nodosum leprosum, a Type 2 leprosy reaction, and explaining its immune-complex-mediated mechanism received 5.375. All three experts agreed with this ranking.
- Prolonged benzodiazepine use for insomnia. “Sleep-Specific Risk Prioritization” requires the response to prioritize rebound insomnia after discontinuation rather than only listing general cognitive adverse effects. “Iatrogenic Risk Framing” distinguishes adverse effects during continued use from the treatment-induced worsening of the original symptom after withdrawal. The negative rubrics “Misplaced Clinical Prioritization” and “Failure to Recognize Treatment-Induced Symptom Escalation” penalize treating anterograde amne-

Table 10: Hyperparameter analysis on ResearchQA, HealthBench, and DRB.

B	G	G'	No-tool Steps	Tool-use Steps	ResearchQA	HealthBench	DRB
24	6	4	600	700	72.8	50.8	41.7
40	7	5	600	700	74.5	53.1	44.3
48	8	6	600	700	76.9	54.4	45.9
64	10	8	600	700	76.9	54.5	45.5
48	8	6	400	700	74.2	52.7	43.6
48	8	6	800	700	72.3	51.4	42.1

sia or daytime sedation as the most important long-term outcome in this context. A response prioritizing anterograde amnesia received 2.0, whereas a response emphasizing tolerance, dependence, withdrawal, and rebound insomnia received 5.5. All three experts agreed with this ranking.

rather than as an attempt to outperform industry systems that use larger or closed-source models.

K Statistical Evidence

We additionally report the mean and standard deviation over three random seeds (42, 43, and 44) for the experiments on stable synthetic-task difficulty, sustainable evolution, and the indispensability of each module.

Stable synthetic-task difficulty. Table 11 reports the synthetic-task rewards at different training steps with or without proposer evolution, indicating that the proposer with evolution continues to generate more challenging tasks for the evolving solver.

Sustainable evolution. As shown in Table 12, the average performance of HOTE across HealthBench, ResearchQA, and DRB continues to improve from 1200 to 1500 total training steps, with small standard deviations across the three seeds.

Indispensability of each module. Table 13 reports the results after 600 steps of no-tool training over three seeds. HOTE consistently outperforms the variants without proposer evolution or judge evolution on all three benchmarks.

L The DeepResearch-Bench Leaderboard

For completeness, Table 14 reports the five highest-scoring systems on the pre-switch official DeepResearch-Bench leaderboard (Du et al., 2025), which is retained as the *Gemini-2.5 Eval* tab.²

HOTE is positioned as a training method that **evolutionarily overcomes the limitations of a fixed RL training set for a given base model**,

²<https://huggingface.co/spaces/muset-ai/DeepResearch-Bench-Leaderboard>, accessed August 30, 2026.

Table 11: Synthetic-task rewards with and without proposer evolution over three seeds.

Setting	100 steps	150 steps	200 steps	250 steps	300 steps
w/ proposer evolution	2.83 ± 0.30	3.07 ± 0.25	4.11 ± 0.05	3.98 ± 0.18	2.10 ± 0.27
w/o proposer evolution	4.87 ± 0.11	3.95 ± 0.36	3.00 ± 0.16	5.19 ± 0.04	5.45 ± 0.11

Table 12: Average performance across three benchmarks during extended HOTE training.

Method	1200 steps	1300 steps	1400 steps	1500 steps
HOTE	58.0 ± 0.1	59.0 ± 0.2	59.5 ± 0.2	59.9 ± 0.1

Table 13: Performance after 600 steps of no-tool training.

Method	HealthBench	ResearchQA	DRB
HOTE	65.1 ± 0.0	66.3 ± 0.2	51.5 ± 0.0
w/o proposer evolution	62.2 ± 0.2	64.2 ± 0.1	50.6 ± 0.1
w/o judge evolution	61.9 ± 0.0	65.3 ± 0.2	50.7 ± 0.1

Table 14: The five highest-scoring systems on the pre-switch official DeepResearch-Bench leaderboard (the *Gemini-2.5 Eval* tab). The openness information follows the official leaderboard.

Rank	System	Overall	Comp.	Insight	Instruction	Readability	Openness
1	DuMate/Qianfan-DeepResearch	58.03	59.48	61.48	53.87	54.34	Proprietary
2	ZTE-Nebula-DeepResearch	57.27	58.37	59.76	54.06	54.66	Proprietary
3	iFlow-Researcher	57.08	58.24	59.74	53.24	55.05	Undisclosed (TBD)
4	Zhipu Deep Research	57.06	58.15	60.14	53.47	53.88	Proprietary
5	Xiaoyi DeepResearch 6.0	57.00	58.58	59.38	53.58	53.99	Proprietary

Algorithm 1 Dual-mode Hybrid Training Strategy for HOTE

Require: Solver π_{θ_s} , Proposer π_{θ_p} , Judge π_{θ_j} .

Require: Training dataset $\mathcal{D}_{\text{train}}$.

Require: Hyperparameters: Batch size B , Group size G , Number of diverse proposing groups N .

```
1: Initialize: Set initial synthetic tasks  $\mathcal{D}_{\text{syn}} = \emptyset$ .
2: while not converged do
3:   // 1. Hybrid Data Preparation
4:   Sample real tasks  $\mathcal{D}_{\text{real}}$  of size  $B/2$  from  $\mathcal{D}_{\text{train}}$ .
5:   Construct current batch  $\mathcal{S} \leftarrow \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{syn}}$ .
6:   // 2. Hybrid Mode Assignment
7:   Randomly assign inference mode  $m \in \{\text{tool-use}, \text{no-tool}\}$  to each task in  $\mathcal{S}$  (50% each).
8:   // 3. Solver Rollout
9:   For each task  $s_0 \in \mathcal{S}$ , sample  $G$  responses  $\{o_i\}_{i=1}^G \sim \pi_{\theta_s}(\cdot \mid s_0)$  under assigned mode  $m$ .
10:  // 4. Judge Evolution & Evaluation
11:  for each task  $s_0 \in \mathcal{S}$  do
12:    Update active rubrics:  $\mathcal{R}_{s_0}^{\text{active}} \leftarrow \text{Update}_{\pi_{\theta_j}}(s_0, \{o_i\}_{i=1}^G, \mathcal{R}_{s_0}^{\text{active}})$  (Equation 3).
13:    Calculate rewards  $r_i$  for each response  $o_i$  using  $\mathcal{R}_{s_0}$  (Equation 2).
14:  end for
15:  Collect assessments  $\mathcal{A}$  containing all rubrics and rewards.
16:  Generate meta rubrics  $\mathcal{R}^{\text{meta}}$  summarizing weaknesses from  $\mathcal{A}$ ,  $\{o_i\}_{i=1}^G$  and  $\mathcal{R}_{s_0}$  (Equation 4).
17:  // 5. Solver Evolution
18:  Update solver parameters  $\theta_s$  via GRPO (Equation 1) using rewards  $\{r_i\}$ .
19:  // 6. Proposer Evolution
20:  if  $\mathcal{D}_{\text{syn}} \neq \emptyset$  then
21:    Calculate proposer rewards  $\{r_i^p\}_{i=1}^{G'}$  for tasks in  $\mathcal{D}_{\text{syn}}$  (Eq. 5).
22:    Update proposer parameters  $\theta_p$  via GRPO using rewards  $\{r_i^p\}_{i=1}^{G'}$ .
23:  end if
24:  // 7. Diverse Proposing (Next Step Synthetic Data)
25:  Sample  $N$  combinations of tasks and corresponding assessments from  $\mathcal{S}$ .
26:  Proposer generates new synthetic tasks  $\mathcal{D}'_{\text{syn}} = \{o_i^p\}_{i=1}^{G'}$  conditioned on combinations and  $\mathcal{R}^{\text{meta}}$ .
27:  Update  $\mathcal{D}_{\text{syn}} \leftarrow \mathcal{D}'_{\text{syn}}$  for the next iteration.
28: end while
```

Can you answer the question from a patient about a medical condition or concern they have: eacs integrase regimens naive hiv recommended intervals (Part 1)

DR Tulu-8B

Bottom line for patients starting an integrase inhibitor (INSTI) regimen (ART/u2011naive)-n- First visit: get a full baseline labs panel (CBC, glucose/metabolic panel including liver enzymes and creatinine, lipids), a pregnancy test for women of childbearing potential, an HIV RNA (viral load), and a CD4 count. If you have risk factors for integrase resistance (e.g., prior exposure to an INSTI for PrEP/PEP or a partner on INSTI), also get an HIV/u20111 integrase genotype to guide regimen selection. <cite id="9213f5a9-0">Standard genotypic resistance testing is recommended to guide ARV choice; if integrase resistance is a concern (e.g., transmitted exposure or prior INSTI), ensure the genotypic test includes the integrase gene, and consider starting ART per "What to Start" while awaiting results.</cite> <cite id="fc70440c-0">Baseline evaluation should include key labs before ART and selected tests before and after initiation or modification to assess virologic and immunologic efficacy and monitor for drug-related toxicities.</cite> <cite id="fc70440c-1">Clinical guidelines provide a general laboratory monitoring schedule before and after ART initiation, detailing when specific tests (e.g., HIV RNA, CD4, and safety labs) should occur.</cite> <cite id="fc70440c-3">Initial assessment goals include confirming diagnosis, assessing immune suppression, guiding ART, and screening for comorbidities, with a recommended CD4 count and quantitative HIV RNA at baseline.</cite> <cite id="87582469-2">Baseline labs include CBC, metabolic panel (including AST/ALT), lipids, pregnancy test when applicable, HIV RNA, and CD4 cell count.</cite> \n\n Early virologic checks after starting an INSTI: check HIV RNA again in the first 6/u20138 weeks to confirm early viral suppression; then check again at 3/u20136 months to confirm durable suppression. This applies to ART/u2011naive adults/adolescents beginning a modern regimen (most of which are integrase/u2011based). <cite id="9213f5a9-3">Guidelines provide a laboratory testing schedule that includes postu2011ART initiation follow/u2011up with HIV RNA and immunologic monitoring.</cite> <cite id="d7698879-2">Monitoring guidance recommends viral load testing 4/u20136 weeks after starting/switching ART and then 3/u20136 months for ART/u2011naive patients.</cite> <cite id="84504351-4">Postu2011ART visit within four weeks is recommended for most patients, with close follow/u2011up to assess early virologic response.</cite> \n\n Once you are consistently suppressed on an INSTI regimen: if viral load remains undetectable and there are no side effects or clinical changes, many clinicians check viral load yearly and use event/u2011driven or CD4/u2011guided monitoring (e.g., CD4 only if not already undetectable, or CD4 if clinically indicated). Routine CD4 checks every 6/u201312 months are often used for stable, well/u2011suppressed patients when viral load is confirmed undetectable; maintain close clinical and medication/u2011adherence counseling. <cite id="9213f5a9-3">Guidelines allow de/u2011escalation or less frequent monitoring for people with HIV on effective ART and sustained suppression, with tests repeated based on stability and clinical context.</cite> <cite id="84504351-4">After a suppressed early viral load, most patients can be followed in 3/u20134 months, then extended intervals if stable; many stable, suppressed patients are monitored less frequently (e.g., annual viral load and event/u2011driven clinical checks).</cite> <cite id="fc70440c-7">For people with HIV on effective ART, primary care guidance emphasizes routine follow/u2011up aligned with suppression status and clinical stability (often including annual laboratory reassessment and targeted testing if changes occur).</cite> \n\n Integrase/u2011specific considerations at start: if there is concern about transmitted integrase resistance (e.g., high/u2011risk sexual networks, partner on INSTI, prior INSTI exposure), obtain an integrase genotype (and, in some settings, phenotype) to confirm susceptibility before committing to an INSTI; an INSTI/u2011displacing regimen can be used while waiting for results. This helps personalize regimen choice and can avoid resistance/u2011selected failure. <cite id="9213f5a9-0">If INSTI resistance is a concern due to potential transmitted resistance or prior INSTI use (PrEP/PEP), ensure genotypic testing includes the integrase gene; guidance on "What to Start" can be consulted while results are pending.</cite> <cite id="fc70440c-0">Resistance testing is intended to guide ARV regimen selection, and integrase testing should be performed when resistance is suspected.</cite> \n\n What to do

HOTE-8B

Summary\n\n- EACS guidance emphasizes individualized care for patients starting INSTI-based regimens, with standardized baseline workup, close early virologic monitoring, and then less frequent follow-up if suppressed, alongside opportunistic infection (OI) prophylaxis and co-infection management (HCV, TB). <cite id="52c35b6f-0">INSTIs are preferred for initial ART in most patients because of potency and high resistance barrier, but clinicians must account for patient factors (e.g., childbearing potential, comorbidities) when selecting regimens and monitoring plans.</cite> <cite id="52c35b6f-3">First-line effectiveness across modern INSTIs (DTG, EVG/c, RAL) is broadly comparable in real-world cohorts, supporting regimen choice based on availability, safety fit, and monitoring capacity rather than major efficacy differences.</cite> \n\n- Virologic check points are critical in the first 24/u201348 weeks; persistent suppression (e.g., below 200/u20131,000 copies/mL thresholds) permits longer intervals, whereas virologic failure requires early investigation for resistance and switch. <cite id="52c35b6f-3">Real-world failure definitions commonly use two consecutive viral loads /u2265200 copies/mL at 24 weeks or a single /u22651,000 copies/mL while on an INSTI, with 6 months and 12 months as key evaluation windows.</cite> \n\n- INSTI-specific safety (lymphadenopathy, hypersensitivity, hepatotoxicity) should be monitored alongside standard CBC and liver enzymes; adjust intervals if abnormalities occur or persist. Resistance testing is advised when INSTI failure is suspected to guide a switch. <cite id="52c35b6f-1">Resistance to first/u2011generation INSTIs (RAL/EVG) is frequent in failure cohorts, and genotype can guide switching to a second-generation INSTI (e.g., DTG).</cite> \n\nBelow is a practical, EACS-aligned monitoring schedule for treatment/u2011naive adults starting an INSTI-based regimen (e.g., dolutegravir with TFC/FTC or raltegravir/abacavir/lamivudine depending on local formulary and resistance testing), including what to check, typical thresholds, and visit intervals. If you would like, I can tailor this to your partner's regimen, co/u2011infections (e.g., HCV, TB), and pregnancy/LPI status.\n\n# Baseline workup (visit 0)\n\n- HIV lab panel: viral load; CD4+ count and CD4% plus CD4/CD8 ratio; full blood count (with retics); comprehensive metabolic panel (creatinine, GFR, urea, electrolytes, liver enzymes ALT/AST, bilirubin, alkaline phosphatase); lipids; fasting glucose/HbA1c; urinalysis; tuberculosis risk assessment and screening in high/u2011risk patients; opportunistic infections risk assessment (e.g., Pneumocystis jirovecii, Toxoplasma gondii in endemic settings); syphilis/HCV screening regardless of risk; hepatitis B surface antigen/quantitative HBV DNA if at risk. <cite id="52c35b6f-0">Choice of INSTI-based first regimen should be individualized with attention to comorbidities and special situations (e.g., pregnancy), implying robust baseline organ function assessment.</cite> \n\n- Medication review and adherence plan. Discuss OI prophylaxis history (e.g., cotrimoxazole); confirm no contraindication to the chosen INSTI (e.g., hypersensitivity); document childbearing potential and contraceptive counseling if relevant. <cite id="52c35b6f-0">Clinical factors (including childbearing potential) influence INSTI prescribing and need to be considered alongside monitoring.</cite> \n\n- INSTI-specific safety baseline: note baseline lymphadenopathy (/u22653 nodes, >1 cm unless congenital) and any hepatocellular adenoma/liver disease history; start routine ALT/AST at baseline and periodically (see below). <cite id="52c35b6f-3">Adverse events differ across INSTIs in real-world cohorts/u2014monitoring supports early detection and management.</cite> \n\n# Initiation and first 4 weeks\n\n- Medication initiation, adherence support, side-effect diary, and symptom diary (GI, hepatic, skin).\n\n- Return in 2/u20134 weeks (visit 1): assess tolerability; repeat full blood count (to detect anemia/ thrombocytopenia); brief clinical exam: spot viral load only if clinically indicated (e.g., symptoms, adherence concerns). <cite id="52c35b6f-0">Early persistence is typically higher with INSTI-based regimens; regular early supportive visits help sustain adherence while detecting early issues.</cite> \n\n# 4/u20138 weeks (2/u00d7 viral load if blips)\n\n- Draw viral load. If suppressed (<200 copies/mL) with stable labs and no symptoms, continue on current interval (often 8/u201312 weeks if stable); if elevated (/u2265200/u20131,000 copies/mL) without other explanations, investigate adherence, interactions, and consider early resistance testing;

Figure 7: Case 1 (Part 1)

Can you answer the question from a patient about a medical condition or concern they have: eacs integrase regimens naive hiv recommended intervals (Part 2)

if viral load is not suppressed or becomes detectable: Re 2011 check viral load in 4 weeks to exclude early window period or lab variability. If still not suppressed, assess adherence, medication toxicity, and clinical status; consider resistance testing (including integrase) and switching to an alternative regimen if resistance is found or adherence is optimal but suboptimal early response persists. Monitoring includes repeat viral load assessments to document virologic efficacy and identify resistance; if suppression is not achieved, evaluate adherence and consider resistance testing and regimen adjustment. When initiating or changing ART, clinicians should anticipate need for viral load and resistance testing to evaluate response and guide modifications. WHO/global context and corroboration: WHO's recent policies emphasize initiating ART as soon as possible and maintaining suppressions with standard monitoring (HIV RNA when clinically indicated and at defined time points, clinical staging, and CD4 checks when CD4 low or clinically indicated). These align with the US/EU practice of frequent early monitoring after initiation and then less frequent checks if undetectable and stable. WHO guidance has increasingly emphasized early initiation and structured monitoring (including viral load and immunologic assessment); while thresholds and schedules vary by setting, the core principle is treating as soon as possible and monitoring response. Monitoring guidance in major societies remains stable, recommending viral load 4 weeks after starting or switching ART, then 3 months for ART-naïve cohorts. A simple monthly checklist to share with your patient: Days 0-201314 after first dose: know your regimen, side effects to report, and baseline labs and viral load results. 4 weeks post start: come in for the first viral load check to confirm suppression. If not suppressed, recheck in 2 weeks and evaluate adherence and resistance testing. An initial follow-up within four weeks of starting ART is advised to monitor early response and address issues promptly. Repeat viral load assessments are used to confirm response and detect early resistance. In 3 months post start: second viral load check to confirm durable suppression; if suppressed, many patients can extend to annual viral load unless a clinician advises otherwise. CD4 monitoring can be tailored: if CD4 was not very low and you are otherwise stable, a single CD4 at baseline and one in 12 months may suffice, with rapid CD4 checks triggered by illness or new symptoms. Guidelines provide a structured laboratory follow-up schedule and allow less frequent monitoring for stable, suppressed patients. Stable suppressors can be followed in 3 months, with intervals extended if adherence and clinical status remain favorable. Primary care guidance emphasizes individualized, event-driven monitoring (e.g., annual reassessment and testing as needed). Ongoing: keep the prescribing clinician informed of side effects (e.g., depression/suicidality with dolutegravir, anemia/crystalluria with raltegravir, liver enzyme elevations), and use contraception appropriately or confirm pregnancy plans; women of childbearing age should have a pregnancy test and continue periodic counseling. Monitoring includes assessment of drug toxicities and clinical status alongside virologic and immunologic endpoints. Safety labs and targeted assessments are part of routine ART follow-up. Why integrase regimens today? Modern guidelines start most ART-naïve adults/adolescents on a single-tablet regimen, usually containing an integrase inhibitor, reflecting high efficacy and tolerability; these regimens simplify adherence and tend to achieve rapid and durable viral suppression. Recent U.S. guideline updates emphasize that for most people with HIV, initial regimens composed of an integrase strand transfer inhibitor are preferred as first-line therapy for ART-naïve patients. Most available first-line regimens are integrase-based, with strong evidence for efficacy and safety in naive patients.

adjust interval to biweekly if persistent elevations. Definitions of virologic failure commonly use two consecutive copies/mL at 24 weeks or a single 1,000 copies/mL to prompt action. 8 weeks Viral load check. If suppressed: proceed to 6 monthly viral load if stable; continue baseline safety labs (CBC, liver enzymes, lipids, renal); review symptoms, contraception, and OI prophylaxis. If elevated, evaluate adherence, drug interactions, clinical symptoms; consider INSTI resistance testing and prepare to switch if confirmed or strongly suspected. Integrase failure often involves resistance to first-line INSTIs; modern genotypes guide second-line INSTI choice. 3 months (anchor visit) Viral load (or use 24 weeks as a key decision point per many programs). If suppressed and otherwise stable, extend viral load to 12 months; repeat full blood count, liver enzymes, lipids, renal profile; TB/HCV follow-up (see below); discuss sexual health and STI screening; reinforce adherence and nutrition. Many cohorts use 6 months as an intermediate check, consistent with the failure rule: two consecutive 1,000 copies/mL at 24 weeks. 12 months Viral load check. If persistent suppression, consider extending to annual viral load in the absence of alerts (symptoms, adherence problems, or lab abnormalities), otherwise repeat the full baseline monitoring panel and continue standard OI prophylaxis and co-infection management. With modern INSTIs, sustained suppression is common by 12 months in real-world settings; maintenance monitoring then tapers unless issues arise. Ongoing maintenance (years 1+) Annual viral load and comprehensive HIV lab panel (CBC, liver enzymes, lipids, creatinine/urinalysis) if stable on suppressive therapy; more frequent if risk factors change (HCV/TB diagnosis, heavy alcohol use, hepatotoxicity history, or pregnancy). Long-term effectiveness and immune recovery are typical with INSTI-based regimens; routine surveillance ensures durability and detects late adverse events. Co-infections and special situations HCV: screen all adults; treat per liver specialists; on INSTI-based ART, viral load and genotype-guided treatment are feasible (e.g., DAA regimens). Intensify liver enzyme monitoring during and after HCV treatment due to hepatotoxicity risk; schedule more frequent ALT/AST during HCV therapy and for 6 weeks post-treatment. Real-world studies show INSTI regimens achieve high immunological and virologic effectiveness; integrating HCV care on INSTI background is feasible with careful hepatotoxicity monitoring. TB risk: in high TB burden settings, consider TST/IGRA at ART initiation; if latent TB identified, initiate IPT per national TB guidelines; monitor symptoms and chest X-ray per local TB protocols. During active TB, cover with standard anti-TB regimen plus an INSTI regimen with close clinical and lab monitoring. EACS notes that clinical context (e.g., TB risk) shapes initial ART choice and monitoring intensity; early detection and management of OIs are integral to outcomes. Pregnancy and long-term prophylaxis (LTOP)/exclusive post-exposure (EPE) for women who have sex with women: confirm pregnancy as soon as possible; continue ART and, if not already on LTOP/EPE, initiate per gynecology/STI guidance; monitor viral load and CBC monthly during pregnancy and immediately postpartum; more frequent viral load in the third trimester if the regimen does not include zidovudine. Childbearing potential and pregnancy are key considerations in INSTI prescribing and require tailored monitoring and counseling. Resistance testing and switching if failure: Consider resistance testing when early INSTI failure is detected (e.g., 1,000 copies/mL at 24 weeks, or 1,000 copies/mL at any time) to identify INSTI class resistance or problematic mutations, guiding switch to a different INSTI or to a non-INSTI backbone (e.g., dolutegravir/ritonavir or lamivudine/abacavir-based regimens). High rates of resistance to first-generation INSTIs on failure genotypes support genotype-guided selection of second-line INSTIs (e.g., replacing raltegravir with dolutegravir in selected third-line contexts; the principle applies to early failures with INSTI-specific mutations).

Figure 8: Case 1 (Part 2)

Can you answer the question from a patient about a medical condition or concern they have: eacs integrase regimens naive hiv recommended intervals (Part 3)

When to escalate
Not suppressed by 3 months despite good adherence: discuss resistance testing (including integrase), consider a regimen swap (e.g., dolutegravir after prior raltegravir resistance, or an INSTI displacing backbone), and address comorbidities and pharmacology interactions. ^{9213f5a9-0}When early suppression is not achieved, reassess adherence, safety labs, and perform resistance testing with consideration of switching ART.
Become detectable after suppression: promptly evaluate adherence and symptoms, recheck viral load, and perform resistance testing if viral load remains elevated or adherence is optimal; switch if resistance is confirmed. ^{9213f5a9-0}Loss of suppression necessitates closer interval viral load checks, adherence counseling, and potential regimen change based on resistance results.
If you want, I can turn this into a short, printed "monitoring schedule" (dates and labs/tests) tailored to your specific start date, regimen, baseline CD4, and any prior INSTI exposure.

INSTI-specific safety monitoring
Lymphadenopathy: document baseline; re-examine at 4 weeks and at 3 months; if new or enlarging nodes (>1 cm or symptomatic), consider imaging and exclude lymphoma; in congenital lymphadenopathy, monitor with closer liver enzyme checks due to hepatocellular adenoma risk. ^{52c35b6f-3}Lymphadenopathy is a common early phenomenon with INSTIs; structured monitoring helps distinguish benign from malignant processes.
Hepatotoxicity: baseline ALT/AST; then at 4 weeks, 3 months, and as indicated; discontinue if ALT >3x upper limit of normal (ULN) with u80e6/u7ea2/u7d20/u5347/u9ad8 or encephalopathy; avoid hepatotoxic medications; in resolved elevations, consider challenge under hepatology supervision if the specific regimen is preferred. ^{52c35b6f-3}Adverse events including hepatic transaminase elevations have been reported with INSTIs; systematic monitoring enables early mitigation.
Practical tips to reduce lab burden for the patient
Combine labs where feasible: run full blood count + metabolic panel on the same day every 3 months; run a lipid panel at 6 and 12 months, then annually if stable; annual comprehensive metabolic panel may suffice after year 1 if liver enzymes and creatinine have been stable.
Use viral load decision points to cluster visits: align 8 weeks and 24 weeks viral load draws with other routine tests (CBC, liver enzymes, renal) to minimize visits.
Use digital tools for symptom tracking (hepatobiliary symptoms, rash, GI intolerance), prompting earlier clinician review if red flags arise (e.g., jaundice, severe fatigue, dark urine). ^{52c35b6f-0}Because INSTI prescribing considers patient factors and tolerability, proactive symptom monitoring and counseling reduce switching and improve long-term adherence.
Notes on regimen choice and regional availability
Second-generation INSTIs (dolutegravir, bictegravir) are preferred for first-line globally; raltegravir remains an option where second-generation INSTIs are unavailable. Choice should consider drug interactions (e.g., with methotrexate or anticoagulants), pregnancy intentions, and patient preference. ^{52c35b6f-0}Guidelines recommend INSTIs for most naive patients; formulary and comorbidity constraints can shape regimen selection within EACS's preference framework.
If the regimen includes abacavir, perform baseline HLA-B*57:01; if positive, switch to a regimen without abacavir. If lamivudine is used, note that resistance on lamivudine failure is uncommon, but other nucleosides should be considered depending on resistance testing. ^{52c35b6f-3}In salvage and first-line populations, nucleoside backbone selection and INSTI choice both impact virologic and immunologic outcomes; real-world data show broadly similar effectiveness across INSTIs, making avoidance of resistance liabilities (e.g., abacavir hypersensitivity) a key practical concern.
If you'd like, please share:
Exact regimen (e.g., DTG + TFC/FTC or RAL + 3TC/ABC), any co-infections (HCV/TB), pregnancy plans/LTOP, and prior STIs. With these details, I can provide a precise, personalized EACS-aligned monitoring schedule (visit-by-visit with due dates, labs, and viral load thresholds) and a clear plan for what to do if there are viral blips, rash, GI symptoms, or abnormal liver tests. Currently, the guidance above is a general framework that fits most INSTI-based first-line regimens; a tailored plan can incorporate local resistance patterns, waiting times, and your partner's preferences and comorbidities.

Figure 9: Case 1 (Part 3)

Can you answer the question from a patient about a medical condition or concern they have: my eye hurts, one time in the past i had acute angle closure glaucoma (Part 1)

DR Tulu-8B

Key points you should know now'n/- Your history of an acute angle'u201closure glaucoma (AACG) attack raises your risk of another acute attack, especially if you have anatomically narrow angles, a shallow anterior chamber, a thick lens, plateau iris, hyperopia, female sex, older age, or a family history. Prior AACG also increases risk in the fellow eye and can leave signs on exam such as glaukomflecken, peripheral anterior synechiae, or an ovoid pupil that point back to a prior angle closure event. <cite id="a69b43b0-1">General risk factors for acute angle closure glaucoma include advanced age, female sex, strong family history, hypermetropia, and Asian/Eskimo ethnicity; anatomical risk factors include a thick crystalline lens, plateau iris, and narrow iridocorneal angles.</cite> <cite id="a69b43b0-3">Signs of a prior attack to look for include glaukomflecken, iris atrophy, an ovoid pupil, and peripheral anterior synechiae.</cite> <cite id="a69b43b0-4">Not every person with narrow angles and a shallow anterior chamber gets an acute attack, but these features heighten risk and additional parameters should be considered when evaluating predisposition.</cite>n/- High intraocular pressure (IOP) is what causes the severe eye pain typical of an AACG attack: prompt treatment is critical to avoid vision loss. <cite id="3fab0317-0">Only very high intraocular pressure, as seen with acute angle-closure glaucoma, causes pain; outpatient argon laser iridectomy is one approach used for treatment.</cite>n/- Glaucoma is one of the top differentials when a red eye is accompanied by deep, severe pain, photophobia, halos around lights, and/or blurred vision. <cite id="2c80b7ae-4">Deep, intense pain, photophobia, and halo vision are characteristic of corneal abrasions, scleritis, iritis, acute glaucoma, sinusitis, etc.</cite> <cite id="2c80b7ae-7">Persistence of eye pain is a red flag, and painful vision loss can be due to acute angle-closure glaucoma, optic neuritis, uveitis, or corneal issues.</cite> <cite id="2c80b7ae-3">New double vision with red eye suggests orbital pathologies such as orbital cellulitis or thyroid eye disease, and any contact or systemic signs of infection raise red flags.</cite>n/n# Why your past AACG matters now'n/- People who have had AACG typically have a narrow anterior chamber and/or closed iridocorneal angles on gonioscopy; these anatomical features persist and can predispose to future acute attacks, even if years have passed without symptoms. <cite id="a69b43b0-1">Many AACG presentations due to pupillary block occur in people unaware that they have innately narrow iridocorneal angles, making prior ocular examination and gonioscopy important.</cite> <cite id="a69b43b0-3">Diagnostic methods such as the van Herick anterior chamber depth grading, indirect gonioscopy (Shaffer/Spaeth), and noncontact imaging like UBM, anterior segment OCT, Scheimpflug, or Pentacam are used to identify angle closure risk and prior subclinical attacks.</cite>n/- Even after a successful laser iridotomy for AACG, closure can sometimes recur, particularly if the iridotomy is narrow or closed, or if there are underlying predispositions (e.g., plateau iris, shallow chamber). In such cases, surgical iridectomy may be needed to permanently relieve pupillary block. <cite id="3fab0317-2">Recurrence of acute angle-closure glaucoma can occur despite Nd:YAG laser iridotomy; surgical iridectomy can permanently resolve pupillary block in these cases.</cite> <cite id="3fab0317-8">A randomized trial found that laser iridotomy in one eye halved the risk of new angle-closure disease compared to the untreated fellow eye, supporting its preventive role.</cite>n/n# Recognizing an acute AACG attack (urgent red flags)n/nIf you develop any of these, seek emergency care immediately:n/- Severe, unrelenting eye pain that worsens with eye movements or pressure over minutes to hours.n/- Sudden blurred vision, "halos" around lights, or a curtain/gray shadow descending with vision loss.n/- Nausea, vomiting, headache, or elevated pulse rate (systemic stress response to pain and elevated IOP).n/- Eye redness, photophobia, and markedly decreased visual acuity in one eye.n/These combinations imply an AACG attack or another sight'u201threatening ocular emergency. <cite id="3fab0317-0">Only very high IOP as in AACG typically causes pain, underscoring the need for urgent management.</cite> <cite id="2c80b7ae-0">Primary care frameworks for red eye identify moderate pain, photophobia, and decreased visual acuity as red flags warranting urgent referral.</cite> <cite id="2c80b7ae-7">Persistent eye pain and painful vision loss should trigger urgent ophthalmologic evaluation to exclude

HOTE-8B

Key points and immediate next steps/n/n- Acute angle-closure glaucoma (AACG) is a true eye emergency that causes sudden, severe eye pain, often with blurred vision, haloes around lights, frontal headache, and nausea or vomiting; the eye exam typically shows very high intraocular pressure (IOP) and a shallow anterior chamber formed by the pupil blocking the trabecular meshwork, so immediate evaluation and treatment are required to prevent rapid, irreversible vision loss<cite id="58932d20-2">"Acute angle-closure glaucoma presents as a sudden onset of severe unilateral eye pain or a headache associated with blurred vision, rainbow-haloes around lights, and rapid IOP rise due to pupillary block at the trabecular meshwork.</cite><cite id="58932d20-3">"Symptoms of AACG include acute severe pain, blurred vision, haloes, frontal headache, nausea, and vomiting; it is an ophthalmic emergency.</cite><cite id="58932d20-8">"Acute angle-closure glaucoma is an emergency and should be at the top of differentials in painful red eyes due to risk of rapid vision loss.</cite><cite id="58932d20-6">"The florid, symptomatic AAC presentation is due to sudden total occlusion of the trabecular meshwork, producing rapidly rising IOP and vision-threatening complications.</cite></n/n->Because you have had AACG before, your risk of angle closure is increased, particularly as the eye's natural lens changes with age (protein liquefaction and cortical swelling can worsen pupillary block and shallowing). Your eye doctor will assess how far the pupil has formed synechiae (adhesions) to the trabecular meshwork and how open the angle is now, and they will also check for any lens opacities or other structural changes contributing to risk<cite id="58932d20-6">"Angle-closure arises from mechanical closure of the trabecular meshwork and can be precipitated by pupillary block mechanisms; evaluation includes assessing the angle and posterior chamber.</cite><cite id="58932d20-9">"Angle-closure and narrow-angle glaucoma arise from structural narrow angles; symptoms include eye pain, haloes, and blurred vision when angle closure occurs.</cite></n/n->If AACG recurs, IOP is brought down urgently, and definitive treatment is a laser peripheral iridotomy (LPI), a small laser hole made in the periphery of the iris to allow aqueous to flow and prevent future closure. LPI is planned after the acute attack is controlled and has become a standard of care; it is generally safe but has trade-offs (not appropriate during an acute attack and carries a small risk of complications, though overall safety is high in properly selected cases)<cite id="58932d20-6">"Management of AAC includes definitive measures to prevent recurrence; standard care has evolved to include laser iridotomy to prevent future angle closure.</cite><cite id="58932d20-10">"After controlling an acute attack, definitive treatment is laser peripheral iridotomy to create an aqueous passage and prevent future angle closure.</cite><cite id="f7492f1c-1">"In suspected primary angle closure, LPI increases angle width and has a good safety profile (approximately 96% with favorable outcomes in series reported by one group).</cite><cite id="f7492f1c-8">"Long-term outcomes suggest LPI provides substantial benefit, reducing progression compared with no treatment, although absolute rates still show residual risk.</cite><cite id="f7492f1c-2">"Laser iridotomy is usually successful but can fail to fully open the angle in a minority of cases (about one in four in some reports), and timing/technique matter.</cite></n/n->Red flags right now (seek care promptly if present)/n/Although your current pain may be due to a non-glaucomatous cause, given your prior AACG, call your eye doctor or go to an urgent care/ER immediately if you have any of these/n/n- Sudden worsening eye pain with blurred vision, haloes, headache, nausea, or vomiting<cite id="58932d20-2">"Acute AACG is characterized by sudden, severe pain with blurred vision and haloes because of rapid pupillary block and IOP rise.</cite><cite id="58932d20-3">"Urgent history clues include severe pain, haloes, blurred vision, headache, nausea, and vomiting.</cite></n/n->Rapid vision loss, fixed and mid-dilated pupil, or inability to feel the light in the affected eye (could indicate angle closure or optic nerve/retina emergency)<cite id="58932d20-6">"Florid AAC can produce rapid vision loss from high IOP causing optic nerve and retinal toxicity; immediate treatment is required.</cite></n/n->Pupillary light reflex not working in the affected eye (can suggest angle closure or optic nerve damage)<cite id="58932d20-6">"Angle closure can compromise anterior segment perfusion and nerve function

Figure 10: Case 2 (Part 1)

Can you answer the question from a patient about a medical condition or concern they have: my eye hurts, one time in the past i had acute angle closure glaucoma (Part 2)

acute angle-closure glaucoma and other serious causes. Considering other causes of current eye pain Because eye pain has many causes, an experienced clinician will distinguish glaucoma from ocular surface, uveitis, and orbital problems. Key distinctions: Corneal abrasion/foreign body: sharp pain, foreign body sensation, tearing, and often a gritty feeling; halos can also occur if there's associated corneal edema or angle closure. Scleritis/episcleritis: deep aching pain, marked redness, often worse with eye movements; systemic features like fever may suggest inflammation. Uveitis: pain with photophobia, floaters, blurred vision; often worse with light. Acute bacterial conjunctivitis: redness, purulent discharge; pain varies and is usually less severe unless complicated or associated with a foreign body/abrasion. Orbital cellulitis or preseptal cellulitis: redness, swelling, proptosis, restricted eye movement, double vision, fever; ENT/ophthalmology urgently. Glaucoma: deep, boring pain; halos; sudden vision loss; elevated IOP and closed/shallow angles on gonioscopy or imaging. Use the table below as a quick mental checklist to help you communicate with your doctor: Features suggesting glaucoma: halos, abrupt blurred vision, painful visual loss, severe deep ache, positive shallow chamber/narrow angles on prior testing or family history of angle closure. Halo vision and acute glaucoma are tightly linked in red-eye differentials. Patients with prior angle-closure may have glaukomaflecken and peripheral anterior synechiae on exam. Features suggesting corneal/ocular surface: foreign body/grittiness, severe localized epithelial defect seen on exam, rapid relief with lubricating drops (often), minimal systemic illness. Corneal abrasions present with pain, photophobia, and foreign body sensation. Keratitis commonly shows eye pain, redness, blurred vision, and photophobia with corneal staining. Features suggesting inflammation/uveitis: pain worsened by light, floaters, ciliary flush, mid-peripheral iris changes. Iritis presents with pain, photophobia, and often a dilated, nonreactive pupil. Features suggesting infection of lids/orbit: proptosis, restricted movement, double vision, fever; urgent referral. New double vision with red eye suggests orbital pathologies such as orbital cellulitis. Primary angle-closure: Primary angle-closure guidance supports urgent ophthalmology referral when infection or red flags are suspected. What your doctor will check (and why): History and vitals: visual acuity changes, onset and character of pain, photophobia, halos, prior episodes, medications (including pilomotorics), and any systemic illness. GPs are taught to document visual acuity, pain, photophobia, and whether findings are unilateral or bilateral in all red eye presentations. Slit-lamp exam: tear film, corneal abrasion/foreign body, fluorescein staining, anterior chamber reaction, lens changes (cataract thickness), and pupillary response. Anatomical risk features and prior attack signs can be identified with slit-lamp evaluation. Tonometry: measures IOP (often in the 30-35 mmHg range during an acute AACG attack). Gonioscopy/anterior segment imaging: directly visualizes the angle to confirm closure or narrow angles; UBM/OCT/Scheimpflug can quantify chamber depth and angle openness. Gonioscopy and noncontact imaging (UBM/OCT/Scheimpflug, Pentacam) are helpful in identifying angle closure risk and prior subclinical attacks. These modalities help stratify risk and determine whether intervention can prevent progression. Visual field testing (if chronic/ongoing): detects peripheral vision loss from prior glaucoma damage. By the time optic nerve damage and field defects occur, prior attacks have often happened; early detection via imaging is emphasized. Treatment and prevention tailored to your history: Acute AACG attack: Immediate medical and sometimes surgical measures to lower IOP rapidly and relieve pupillary block. Argon laser iridectomy is a mainstay to create an alternative outflow pathway and prevent further angle closure. Argon laser iridectomy is used in the outpatient management of angle-closure glaucoma. If the iridotomy is inadequate or recurrent closure occurs, surgical iridectomy or other procedures may be required. Recurrent AACG after Nd:YAG laser iridotomy may require surgical iridectomy to permanently resolve

acutely; exam findings guide urgency. Severe eye redness with swelling around the eye, proptosis (bulging), diplopia (double vision), or fever these favor an orbital/infectious cause needing antibiotics or surgical drainage decisions. In painful red eye algorithms, acute angle-closure is at the top, but infectious/structural causes (e.g., preseptal/orbital cellulitis) present with significant edema, proptosis, diplopia, and systemic signs; rapid specialist care is needed. Sudden, severe unilateral headache with nausea/vomiting and possible neurological deficits; consider cerebral causes alongside ocular ones inappropriately managed. Neurologic presentations can accompany severe AACG; triage by symptoms including headache and vomiting. What your eye doctor will check and why it matters: Intraocular pressure and corneal thickness: rapid IOP spikes cause pain and can create corneal edema with halos; both IOP and corneal mapping (tonometry and pachymetry) are part of acute assessment. Rapid IOP elevation is central to AACG's symptoms and exam findings (pain, halos) because of pupillary block. Gonioscopy (angle view) to see how far the pupil has stuck to the trabecular meshwork (posterior synechiae) and how open the angle is now. This informs prognosis and whether preventive laser (LPI) is indicated, even if you're not in an acute attack. Gonioscopy is used in angle-closure to assess closure, synechiae, and angle opening for management decisions. Glaucoma guideline frameworks emphasize testing and specialty-referral pathways for suspected angle closure and recurrent/cataract-related risk. Anterior chamber depth and lens status (cataract exam): progressive cataract formation (especially cortical "spunous" changes) increases risk of pupillary block; early cortical changes can be detected with slit-lamp exam and OCT/corneal topography sometimes. Your doctor may recommend closer follow-up or cataract surgery timing based on this risk. Narrow-angle glaucoma and angle-closure risk escalate with structural shallowing and cataract-related pupillary block; cataract surgery can both lower risk and improve visual performance. Visual fields and optic nerve: after a prior AACG, even with controlled IOP, some visual field loss can remain; baseline testing and periodic repeat are part of long-term care in guidelines. Glaucoma guidelines recommend comprehensive assessment (visual function and structural testing) and structured follow-up for patients at elevated risk. Preventive strategies after a prior AACG: Laser peripheral iridotomy (LPI) after acute attack control: LPI creates an aqueous passage in the iris periphery to prevent future angle closure; it is now the definitive treatment of choice and can be planned as an outpatient procedure once the eye is stable after an episode. Once the acute attack is controlled, definitive treatment is laser peripheral iridotomy to provide a small drainage passage and prevent future angle closure. LPI can increase angle width and has a good safety profile, supporting its use in appropriate candidates. Medical "angle-opening" drops during at-risk periods: In eyes that intermittently show high risk (e.g., early morning shallowing with high IOP "dawn phenomenon," or transient pupillary block), clinicians may use miotic agents (pilocarpine ophthalmic) at night to pull the pupil away from the trabecular meshwork; these drops must be used under medical supervision because they can precipitate an attack inappropriately when not indicated. After acute attack control, definitive LPI is standard; still, in selected "angle-closure suspect" eyes, topical miotic therapy may be used at night to increase angle width and reduce risk (with close monitoring to avoid precipitating an attack). Cataract management: Addressing the underlying lens changes that drive pupillary block is central to long-term prevention. In many patients with prior AACG and clinically significant cataract (especially cortical), cataract surgery is timed to reduce angle closure risk; surgery itself can improve angle openness and visual function. Cataract-related pupillary block is a common driver of angle closure; cataract surgery can lower narrow-angle risk and improve vision, with angle-opening benefits often seen in phakic and post-cataract states. Historical consensus emphasizes that definitive surgical intervention (peripheral iridotomy) protects against recurrent episodes and damage from intermittent attacks, and cataract-related risk is a major driver of

Figure 11: Case 2 (Part 2)

Can you answer the question from a patient about a medical condition or concern they have: my eye hurts, one time in the past i had acute angle closure glaucoma (Part 3)

pupillary block.</cite>\n\n Long-term prevention (given your prior AACG):</cite>\n - Consider laser peripheral iridotomy (LPI) in the fellow eye or in angle-closure suspects to widen the angle and reduce risk, especially if you have narrow angles, shallow chamber, hyperopia, or other risk factors. Randomized evidence supports a roughly 50% reduction in new attacks with LPI.</cite>\n - Laser iridotomy in one eye halved the risk of new angle-closure disease compared with the untreated eye in a randomized trial.</cite>\n - In angle-closure suspects, LPI is often indicated when any pupillary block potential is present.</cite>\n - Laser and surgical treatments for angle closure can delay or prevent elevated IOP and progression to chronic angle-closure glaucoma.</cite>\n - Expect that LPI is generally effective but not perfect; success rates vary and can be lower in some populations, and recurrence can occur if the iridotomy becomes obstructed or if underlying risk factors remain. Surgical iridectomy may be curative in cases of recurrent closure after YAG LPI.</cite>\n - Angle-closure risks persist without LPI, and some studies report low but definite future risk of angle-closure glaucoma even after prophylaxis.</cite>\n - Reported success rates for laser iridotomy range from 65% to 137%, with relatively lower success in some Asian populations, and identifying prognostic factors helps guide expectations.</cite>\n - Some patients with prior LPI will present recurrent AACG and require surgical iridectomy to permanently relieve pupillary block.</cite>\n - Be cautious with miotic (pilocarpine) eye drops in angle-closure glaucoma unless an ophthalmologist specifically deems them safe; miotics can worsen pupillary block in narrow angles. Instead, doctors often favor cycloplegic dilating drops to relax the pupil during prophylaxis/evaluation when angle closure risk is present. (General counseling: Discuss all medications, including over-the-counter and supplements, with your doctor and avoid presumed "pilots" without medical discussion.)</cite>\n - Patients known to have innately narrow angles should be referred for gonioscopy and stratification due to risk of acute angle closure; avoid self-medication that could precipitate an attack.</cite>\n\n Practical next steps:</cite>\n - If you have any of the urgent red flags listed above (severe pain, sudden halos/blur with or without vision loss, systemic illness, double vision), go to the nearest emergency department now. Glaucoma and other ocular emergencies need rapid evaluation and treatment to preserve vision.</cite>\n - Severe pain in AACG is due to very high IOP and requires urgent intervention.</cite>\n - Moderate pain, photophobia, and visual changes in red-eye presentations are red flags indicating urgent referral.</cite>\n - Persistent eye pain and painful vision loss should be assessed urgently to exclude serious causes like acute angle-closure glaucoma.</cite>\n\n If not an emergency, make an urgent appointment with an ophthalmologist (often within 1-3 days) to:</cite>\n - Re-assess angle openness (gonioscopy) and anterior segment anatomy (OCT/UBM if indicated).</cite>\n - Review medication safety and adjust preventive measures (e.g., laser iridotomy candidacy).</cite>\n - Consider baseline visual fields and imaging if chronic risk remains.</cite>\n\n Your past AACG means you're a "primary angle-closure suspect" and should be managed proactively.</cite>\n - Primary angle closure glaucoma risk includes prior attacks, narrow angles, and features like peripheral anterior synechiae; early detection methods are available.</cite>\n - Laser iridotomy reduces the risk of new angle-closure events and is a mainstay of prevention in high-risk eyes.</cite>\n\n In the interim, if no red flags are present but you have mild eye irritation/redness without pain or vision changes, use artificial tears as first-line and protect the eye from further irritation. If symptoms evolve toward pain, halos, or vision loss, reclassify immediately as urgent.</cite>\n - Common causes like corneal abrasions and superficial disease cause pain/photophobia and often respond to lubrication and avoidance of trauma.</cite>\n - Primary-care guidance emphasizes recognizing red flags and timely referral over time-sensitive self-treatment for suspected serious causes.</cite>\n\n Key facts to bring to your ophthalmologist:</cite>\n - Tell them about your prior AACG, including age at attack, IOP at presentation, how you were treated, and any family history of	recurrence.</cite>\n - Structured follow-up: Given your history, your doctor may schedule more frequent gonioscopy, tonometry, and field/nerve checks over time, and discuss imaging if there are visual field deficits; guidelines support enhanced monitoring for people with prior angle closure and risk factors.</cite>\n - Glaucoma guideline recommendations include enhanced testing and follow-up for patients with prior angle closure and anatomical risk factors.</cite>\n - When LPI is appropriate:</cite>\n - Indicated: after a confirmed or strongly suspected AACG has been treated to lower IOP; planned as an outpatient laser procedure to permanently relieve angle closure risk; also considered prophylactically in "primary angle closure suspect" eyes before an acute attack occurs.</cite>\n - LPI is performed after IOP control in AACG as definitive therapy to prevent recurrence.</cite>\n - In primary angle closure suspect, LPI increases angle width; safety profiles are favorable in selected cases.</cite>\n - Not appropriate during an acute attack: LPI is performed only after IOP is controlled; during AACG the priority is rapid medical decompression and supportive care to halt vision loss and IOP damage (topical and systemic osmotics, miotics only under supervision, and sometimes systemic analgesia/antiemetics), followed by emergency LPI once stable.</cite>\n - Acute management focuses on rapid IOP reduction; definitive LPI is deferred until after acute control to ensure safety and efficacy.</cite>\n - Laser procedures must be timed appropriately; failure or emergent use during high-risk states can be ineffective or dangerous.</cite>\n - Risks: Like any anterior segment surgery, LPI can cause anterior segment inflammation/cataract progression and, very rarely, malignant glaucoma (a difficult post-surgery angle closure), which may require surgical revision; overall safety is good when applied to appropriate candidates and performed by experienced surgeons (series report high success and safety rates with careful selection).</cite>\n - Safety profiles for LPI can be favorable in selected cases (approximately 96% in series reported), but careful case selection is required.</cite>\n - Historical consensus notes that while LPI prevents recurrent angle closure and protects the eye, risks exist and require careful evaluation and follow-up.</cite>\n - Brochures explain that LPI has known benefits and risks, including cataract progression and rare cases of malignant glaucoma requiring further surgery.</cite>\n\n Other causes of current eye pain and when to seek care:</cite>\n - Not all painful eyes are AACG; other conditions can present with pain and require different treatments:</cite>\n - Dry eye/corneal exposure: persistent pain with gritty sensation, intermittent blurred vision, and minimal vision loss; treated with artificial tears and environmental/temporary lid occlusion.</cite>\n - Eye pain differentials include dry eye and lid/ocular surface disease; a careful red-eye and corneal assessment is needed to exclude serious causes.</cite>\n - Blepharitis or meibomian gland dysfunction: eyelid margins are crusted/irritated; lid scrubs and warm compresses help.</cite>\n - Lid margin disease can cause ocular surface pain and irritation but must be distinguished from more urgent ocular conditions.</cite>\n - Viral/ bacterial conjunctivitis, corneal abrasion/ulcer: red eye with discharge or foreign body history, photophobia; "map-dot-fingerprint" corneal staining suggests herpes simplex keratitis and needs antiviral drops promptly; bacterial keratitis often needs antibiotic drops; all these require clinician evaluation before treatment to avoid masking serious disease.</cite>\n - In red-eye algorithms, infectious causes can cause pain and require specific, promptly initiated therapy; do not self-treat without exam.</cite>\n - Preseptal/orbital cellulitis: rapid swelling around and into the eye, sometimes proptosis, diplopia, fever; urgent ophthalmic and medical-surgical evaluation is needed to differentiate from orbital cellulitis and start antibiotics/surgery if abscess is present.</cite>\n - Signs like toxic appearance, proptosis, diplopia, and fever suggest orbital cellulitis or facial sinus infection; urgent evaluation is required to guide antibiotics and drainage.</cite>\n - Migraine/ tension-type headache or sinus pressure: bilateral pain, worse in morning, nasal congestion/fullness; may improve with analgesics and nasal saline.</cite>\n - still consider an ophthalmic evaluation for optic neuropathy or increased intracranial pressure if visual changes occur.</cite>\n - Headache and nausea
---	--

Figure 12: Case 2 (Part 3)

Can you answer the question from a patient about a medical condition or concern they have: my eye hurts, one time in the past i had acute angle closure glaucoma (Part 4)

glaucoma/angle closure. <cite id="a69b43b0-1">Family history, hyperopia, and anatomical risk factors like narrow angles increase risk and influence management.</cite>\n- Report current symptoms: onset, character, duration, associated photophobia, halos, diplopia, or visual changes, and whether any overu2011counter medications or eye drops made symptoms better/worse. <cite id="2c80b7ae-2">Documenting visual acuity, pain, photophobia, and unilateral/bilateral nature helps organize history.</cite>\n- Share a list of current medications (prescription and OTC, including any eye drops, supplements). Some drops can precipitate angle closure in predisposed patients. <cite id="a69b43b0-1">Patients with narrow angles should be referred and medication histories reviewed to avoid precipitants of acute angle closure.</cite>\n- Ask about baseline testing appropriate for your risk (tonometry, gonioscopy, anterior segment OCT/UBM), prophylactic laser iridotomy considerations, and what to watch for between visits. <cite id="a69b43b0-3">Gonioscopy and imaging help stratify risk of future attacks and guide preventive therapy before optic nerve damage occurs.</cite> <cite id="3fab03f7-8">Evidence supports iridotomy to reduce future angle-closure events in high-risk eyes.</cite>\n\n# Bottom line\n\n- An eye that hurts in someone with a history of AACG should be treated as potentially urgent. AACG can recur and does so more often in people with anatomically narrow angles and other risk factors; prompt recognition and treatment of elevated IOP and pupillary block are essential to preserve vision. <cite id="3fab03f7-0">High IOP is the mechanism of pain in acute angle-closure glaucoma and requires urgent care.</cite> <cite id="a69b43b0-1">Narrow angles and other risk factors increase the chance of recurrence and justify proactive evaluation and prevention.</cite> \n- Bring any new red u2011flag symptoms (sudden halos/blur/vision loss, severe deep pain, double vision, systemic illness) to emergency care. If red flags are absent but you remain at risk, get an urgent ophthalmology evaluation and consider prophylactic laser iridotomy based on your anatomy and other risk factors. <cite id="2c80b7ae-0">Red u2011flag features warrant urgent referral pathways.</cite> <cite id="3fab03f7-8">Laser iridotomy reduces new angle-closure events and is a cornerstone of prevention in appropriate candidates.</cite> \n\nIf you want, I can help you draft a message for your clinician summarizing these points and your symptoms, and walk through a short at u2011home monitoring script to share with your family so you're all on the same page about what to do if symptoms change.

can accompany ocular emergencies; a careful history and exam help separate primary headache from secondary causes.</cite>.\n- Optic neuritis or anterior ischemic optic neuropathy (AION): sudden monocular vision loss with or without eye pain on gaze; see neuro-ophthalmology urgently as these entities often require steroids or other therapies to limit vision loss; prior glaucoma or elevated cup-to-disc can increase concern for AION in older patients<cite id="58932d20-8">Sudden monocular vision loss with pain on gaze should prompt urgent evaluation for optic neuritis; risk factors and ophthalmic exam findings guide next steps.</cite>.\n- Giant cell arteritis (temporal arteritis), especially in those over 50 with new diplopia, headache, jaw claudication; treat urgently with steroids to prevent blindness u2011ask about temple tenderness/redness and visual symptoms; blood tests and imaging can follow if indicated<cite id="58932d20-4">Systemic vasculitis like giant cell arteritis in older patients with new visual symptoms and scalp/temple tenderness is an urgent consideration requiring prompt evaluation and treatment.</cite>.\n\nIn all cases, because you have a prior AACG, report any new or worsening eye pain, visual changes, or symptoms above to an eye care professional promptly for evaluation. If you're unsure whether it's an emergency, go to the closest ED; they can stabilize pain/nausea and arrange rapid ophthalmology review<cite id="58932d20-3">When in doubt with painful red eye and systemic symptoms, seek urgent care; ophthalmology will be consulted for suspected AACG.</cite><cite id="58932d20-8">Prompt specialist involvement is critical in differentials that can cause rapid vision loss or structural compromise.</cite>.\n\n# Practical monitoring tips at home (to discuss with your doctor)\n\n- Note patterns: Does the pain occur in the morning or after long reading/near work? Track any haloes, intermittent blurring, or nausea and how often they happen; morning IOP spikes and transient pupillary block are common triggers in susceptible eyes<cite id="58932d20-2">Pupillary block from near work/light exposure/lens changes can precipitate AACG; recognizing patterns can help prevention and early intervention.</cite>.\n- Vision logs: Document any changes in vision acuity or color vision; color vision loss can be an early sign of optic nerve stress in angle-closure situations<cite id="58932d20-6">Acute angle closure can compromise optic nerve perfusion rapidly; early detection of visual changes is critical.</cite>.\n- Hydration and caffeine: Dehydration and caffeine can thicken and slow aqueous; maintain hydration, limit diuretic/caffeine intake before bed if you have an at-risk morning state; discuss personalized thresholds with your clinician<cite id="58932d20-6">Systemic factors that alter aqueous production or outflow can influence IOP in angle-closure eyes; clinician guidance on modifying factors is part of care.</cite>.\n- Medication review: Some over-the-counter decongestants/sudorifics and even certain asthma inhalers can raise IOP acutely; ask your pharmacist and ophthalmologist to review your medications<cite id="58932d20-6">Medication effects on aqueous dynamics matter in angle-closure eyes; ophthalmic and pharmacy review can reduce avoidable IOP spikes.</cite>.\n- Cataract progression: Schedule annual slit-lamp exams if you've had cataract surgery or have early cortical changes; cortical cataract progression can change angle dynamics and should be discussed with cataract surgeons if you're an appropriate surgical candidate<cite id="58932d20-9">Progressive cortical cataract can worsen angle closure risk; coordinated care between ophthalmology and cataract surgery teams guides timing.</cite>.\n\n# If you are experiencing pain now: what to do\n\n- If you have any AACG red flags (severe unilateral eye pain with haloes/blurred vision, nausea/vomiting, fixed/dilated pupil), go to the nearest emergency department immediately<cite id="58932d20-2">Symptoms of AACG indicate a true ophthalmic emergency requiring rapid decompression.</cite><cite id="58932d20-3">Urgent history and exam are required to diagnose and treat promptly.</cite>.\n- If the pain is persistent but not meeting full AACG red flags, call your ophthalmologist or glaucoma specialist during office hours; describe patterns (time of day, activities that worsen it), associated symptoms, and any visual changes. Arrange an exam to reassess IOP, angles (gonioscopy), and cataract status; baseline visual fields/nerve assessment may be helpful after stabilizing<cite id="58932d20-0">Glaucoma guidelines support structured testing and follow-up, with escalation when risk factors are present.</cite><cite id="58932d20-6">Comprehensive eye exams including angle assessment and cataract evaluation are central to preventing AACG recurrence.</cite>.\n\n# Bottom line\n\n- Your past AACG puts you in a higher-

Figure 13: Case 2 (Part 4)

Can you answer the question from a patient about a medical condition or concern they have: my eye hurts, one time in the past i had acute angle closure glaucoma (Part 5)

-risk group, especially if you have or are developing cortical cataract; rapid IOP spikes can still occur and cause pain, so report any new or worsening symptoms promptly<cite id="58932d20-6">Prior AACG increases long-term risk due to anatomical predisposition to angle closure; vigilant monitoring and timely intervention reduce harm.</cite>.\n- AACG is best prevented with definitive LPI after an attack is controlled, along with proactive management of the lens-related risk (often via cataract surgery) and careful use of "angle-opening" drops only under supervision<cite id="58932d20-10">LPI is definitive therapy after IOP control in AACG to prevent recurrence.</cite><cite id="58932d20-9">Addressing cataract-driven pupillary block reduces recurrence risk and improves vision.</cite>.\n- If you are experiencing current pain, use the red-flag checklist above to decide: go to the ER for AACG-like features; call your doctor for persistent, non-emergent pain or changes; don't delay evaluation, given your history<cite id="58932d20-2">Acute AACG requires immediate medical attention.</cite><cite id="58932d20-8">Prompt evaluation is essential to preserve vision in painful red eye presentations.</cite>.\n\nIf helpful, bring this summary to your next eye care visit so we can review your specific findings and tailor a monitoring and prevention plan (including timing of cataract surgery, LPI candidacy, and home tracking).

Figure 14: Case 2 (Part 5)

System prompt of solver under tool-use mode (Part 1)
You are a research assistant who answers questions through iterative reasoning and research. ## Process - Use <code><think></think></code> tags to show your reasoning at any point. - Use <code><call_tool name="...">query</call_tool></code> when you need information (see tools below). - You can alternate between thinking and searching multiple times. - Only provide <code><answer></answer></code> tags when you have enough information for a complete response. If the problem asks for a specific, short-form answer, you can also put the answer string in the <code>\boxed{}</code> format. - Support every non-trivial claim with retrieved evidence. Wrap the exact claim span in <code><cite id="ID1,ID2">...</cite></code>, where id are snippet IDs from searched results (comma-separated if multiple). Use only returned snippets; never invent IDs. Avoid citing filler text - cite just the factual claim. ## Calling Tools (<code><call_tool name="...">query</call_tool></code>) - You can use the following tools: 1. <code>google_search</code> - Purpose: general web search. - Input via: <code><call_tool name="google_search">your query</call_tool></code> - Output: web search snippets (see SEARCH RESULTS). - Optional parameters - <code>gl</code>: geolocation - <code>hl</code>: host language 2. <code>browse_webpage</code> - Purpose: open a specific URL (typically one returned by <code>google_search</code>) and extract readable page text as snippets. - Input via: <code><call_tool name="browse_webpage">https://example.com/article</call_tool></code> - Output: webpage (see SEARCH RESULTS). 3. <code>snippet_search</code> - Purpose: focused snippet retrieval from scientific papers - Input via: <code><call_tool name="snippet_search">your query</call_tool></code> - Output: snippets from existing papers (see SEARCH RESULTS). - Examples: <code><call_tool name="snippet_search" limit="8" year="2021-2025" fieldsOfStudy="Computer Science, Medicine">large language model retrieval evaluation</call_tool></code>

Figure 15: System prompt of solver under tool-use mode (Part 1)

System prompt of solver under tool-use mode (Part 2)

- Optional parameters

- limit: number of snippets to retrieve
- year: publication year; you can use a single number (e.g., 2024) or a range (e.g., 2022-2025)
- fieldsOfStudy: One or a comma-separated list from: Computer Science, Medicine, Chemistry, Biology, Materials Science, Physics, Geology, Psychology, Art, History, Geography, Sociology, Business, Political Science, Economics, Philosophy, Mathematics, Engineering, Environmental Science, Agricultural and Food Sciences, Education, Law, Linguistics.

Tool Output

- After you issue a tool call, we will execute it and return results wrapped in <tool_output> tags.
- For web search and snippet search, the results appear as: <tool_output><snippet id=UNIQUE_ID>content</snippet>...</tool_output>
- For web browsing, the searched results are represented as <tool_output>content</tool_output>

Answer and Citation Format

- Once you collect all of the necessary information, generate the final answer, and mark your answer with answer tags: <answer></answer>.
- If your answer is short (e.g., a phrase or a number), you can also put the answer string in the \boxed{} format.
- In your answer, wrap the supported text in <cite id="SNIPPET_ID"> ... </cite>. You have to use the exact ID from a returned <snippet id=...>...</snippet>.
- If multiple sources support a passage, use multiple <cite> tags around the relevant clauses/sentences.

- Examples

<cite id="S17">LLMs often hallucinate on long-tail facts.</cite>

<answer>Based on the search results, <cite id="S23">the first Harry Potter movie was released on November 16, 2001.</cite>Therefore, the final answer is \boxed{November 16, 2001}.</answer>

WORKFLOW EXAMPLE

...

REQUIREMENTS

- Think and search iteratively until you have sufficient information
- Only provide the final answer when ready
- Cite all claims from search results using exact snippet IDs

Figure 16: System prompt of solver under tool-use mode (Part 2)

System prompt of solver under no-tool mode

You are a research assistant who answers questions through careful reasoning.

Process

- Use <think></think> tags to show your reasoning at any point.
- Think through the problem step by step, considering different aspects and perspectives.
- Only provide <answer></answer> tags when you have formulated a complete response.
- If the problem asks for a specific, short-form answer, you can also put the answer string in the \boxed{} format.
- Base your response on your knowledge and reasoning abilities.

Answer Format

- Once you have reasoned through the question, generate the final answer, and mark your answer with answer tags: <answer></answer>.
- If your answer is short (e.g., a phrase or a number), you can also put the answer string in the \boxed{} format.
- Structure your answer clearly and provide supporting reasoning.

WORKFLOW EXAMPLE

...

REQUIREMENTS

- Think through problems carefully before providing your answer
- Structure your response clearly
- Provide well-reasoned explanations

Figure 17: System prompt of solver under no-tool mode

System prompt of proposer under tool-use mode (Part 1)
You are an expert question generation assistant specializing in creating challenging research questions. Your goal is to analyze meta rubrics, previous questions, and use research tools to generate novel, difficult questions that target specific quality dimensions. ## Process - Use <think></think> tags to show your reasoning at any point. - Use <call_tool name="...">query</call_tool> when you need research information to generate better questions (see tools below). - You can alternate between thinking and using tools multiple times. - Only provide <answer></answer> tags when you have generated a complete, challenging question that targets the quality dimensions identified in the meta rubrics. ## Understanding Meta Rubrics - Meta Rubrics: Quality dimensions derived from analyzing previous questions. Each rubric identifies a specific capability area with an associated weight. - Weight Interpretation: - Weight = 1.0: Indicates an underdeveloped strength - shows some capability but needs improvement - Weight = -1.0: Indicates a clear weakness - consistently struggles with this capability - Relationship to Previous Questions: Meta rubrics are systematically extracted from patterns observed in previous questions. They represent both areas that need improvement (negative weights) and areas that show promise but fall short of optimal performance (positive weights). ## Calling Tools (<call_tool name="...">query</call_tool>) - You can use the following tools: 1. google_search - Purpose: general web search. - Input via: <call_tool name="google_search">your query</call_tool> - Output: web search snippets (see SEARCH RESULTS). - Optional parameters - gl: geolocation - hl: host language 2. snippet_search - Purpose: focused snippet retrieval from scientific papers - Input via: <call_tool name="snippet_search">your query</call_tool> - Output: snippets from existing papers (see SEARCH RESULTS). - Examples: <call_tool name="snippet_search" limit="8" year="2021-2025" fieldsOfStudy="Computer Science, Medicine">large language model retrieval evaluation</call_tool> - Optional parameters - limit: number of snippets to retrieve - year: publication year; you can use a single number (e.g., 2024) or a range (e.g., 2022-2025) - fieldsOfStudy: One or a comma-separated list from: Computer Science, Medicine, Chemistry, Biology, Materials Science, Physics, Geology, Psychology, Art, History, Geography, Sociology, Business, Political Science, Economics, Philosophy, Mathematics, Engineering, Environmental Science, Agricultural and Food Sciences, Education, Law, Linguistics.

Figure 18: System prompt of proposer under tool-use mode (Part 1)

System prompt of proposer under tool-use mode (Part 2)
3. browse_webpage - Purpose: open a specific URL (typically one returned by google_search or snippet_search) and extract readable page text as snippets. - Input via: <call_tool name="browse_webpage">https://example.com/article</call_tool> - Output: webpage (see SEARCH RESULTS). ## Tool Output - After you issue a tool call, we will execute it and return results wrapped in <tool_output> tags. - For snippet search, the results appear as: <tool_output><snippet id=UNIQUE_ID>content</snippet>...</tool_output> - For web browsing, the searched results are represented as <tool_output>content</tool_output> ## Question Generation Focus - Your primary goal is question generation, not answering research questions. - Use tools to: 1. Understand what makes certain problem types challenging 2. Research current research frontiers and unsolved problems 3. Find domains requiring skills that match both the weaknesses (negative weights) and underdeveloped strengths (positive weights) in meta rubrics 4. Identify problem structures that combine multiple weak areas and underdeveloped capabilities - Generate questions that are: - Challenging but solvable with correct reasoning - Different in scope and complexity from previous questions - Targeted at systematic weaknesses (negative weight rubrics) and underdeveloped strengths (positive weight rubrics) - Requiring precise execution without room for loose interpretation - Similar in form and length to previous questions: Match the style, brevity, and naturalness of previous questions rather than creating overly structured or verbose prompts ## Strategic Research - When using tools, focus research questions on understanding the problem domain and finding appropriate challenge areas rather than seeking direct answers to the target question. ## Answer Format - Once you have analyzed the meta rubrics, previous questions, and any research findings, generate your final question in: <answer>generated question here</answer> ## WORKFLOW EXAMPLE ... ## Requirements - Focus on generating questions that exploit both weaknesses (negative weights) and underdeveloped strengths (positive weights) identified in meta rubrics - Use tools to research challenging problem domains, not to answer questions - Ensure generated questions are different from previous questions - Generate questions that match the form and length of previous questions: Create brief, natural, concise prompts similar to real user questions rather than overly structured or verbose challenges - Only output the final question when ready

Figure 19: System prompt of proposer under tool-use mode (Part 2)

User prompt of proposer under tool-use mode
Meta Rubrics {rubrics_block} **Previous Questions** {prev_queries_block} **Task** Your goal is to generate a new question that targets specific quality dimensions as identified by "meta rubrics". Focus on question generation; do not attempt to answer the question. **Requirements** - Generate a question that is significantly different from previous questions while targeting both weaknesses (negative weights) and underdeveloped strengths (positive weights) described in the meta rubrics. - The question should be challenging and precise, focusing on the identified quality dimensions. - Generate questions that match the form and length of previous questions: Create brief, natural, concise prompts similar to real user questions rather than overly structured or verbose challenges. - Use tools to search from the internet to ensure the reasonableness and challenge of the generated question. For the given meta rubrics and previous questions, please think and search from the internet iteratively to generate a brief but challenging question that is similar in form and length to previous questions. You should use tools to find multiple sources in order to find the needed evidence for the reasonableness and challenge of the generated question. You should provide the final generated question in following format: <answer>generated question here</answer>.

Figure 20: User prompt of proposer under tool-use mode

System prompt of proposer under no-tool mode
You are an expert question generation assistant specializing in creating challenging research questions. Your goal is to analyze meta rubrics and previous questions to generate novel, difficult questions that target specific quality dimensions. ## Process - Use <think></think> tags to show your reasoning at any point. - Think through the problem step by step, considering different aspects and perspectives. - Only provide <answer></answer> tags when you have generated a complete, challenging question that targets the quality dimensions identified in the meta rubrics. ## Understanding Meta Rubrics - Meta Rubrics: Quality dimensions derived from analyzing previous questions. Each rubric identifies a specific capability area with an associated weight. - Weight Interpretation: - Weight = 1.0: Indicates an underdeveloped strength - shows some capability but needs improvement - Weight = -1.0: Indicates a clear weakness - consistently struggles with this capability - Relationship to Previous Questions: Meta rubrics are systematically extracted from patterns observed in previous questions. They represent both areas that need improvement (negative weights) and areas that show promise but fall short of optimal performance (positive weights). ## Question Generation Focus - Your primary goal is question generation, not answering research questions. - Think carefully about: 1. What makes certain problem types challenging 2. Current research frontiers and unsolved problems 3. Domains requiring skills that match both the weaknesses (negative weights) and underdeveloped strengths (positive weights) in meta rubrics 4. Problem structures that combine multiple weak areas and underdeveloped capabilities - Generate questions that are: - Challenging but solvable with correct reasoning - Different in scope and complexity from previous questions - Targeted at systematic weaknesses (negative weight rubrics) and underdeveloped strengths (positive weight rubrics) - Requiring precise execution without room for loose interpretation - Similar in form and length to previous questions: Match the style, brevity, and naturalness of previous questions rather than creating overly structured or verbose prompts ## Answer Format - Once you have analyzed the meta rubrics and previous questions, generate your final question in: <answer>generated question here</answer> ## WORKFLOW EXAMPLE ... ## Requirements - Focus on generating questions that exploit both weaknesses (negative weights) and underdeveloped strengths (positive weights) identified in meta rubrics - Think carefully about challenging problem domains - Ensure generated questions are different from previous questions - Generate questions that match the form and length of previous questions: Create brief, natural, concise prompts similar to real user questions rather than overly structured or verbose challenges - Only output the final question when ready

Figure 21: System prompt of proposer under no-tool mode

User prompt of proposer under no-tool mode
Meta Rubrics {rubrics_block} **Previous Questions** {prev_queries_block} **Task** Your goal is to generate a new question that targets specific quality dimensions as identified by "meta rubrics". Focus on question generation; do not attempt to answer the question. **Requirements** - Generate a question that is significantly different from previous questions while targeting both weaknesses (negative weights) and underdeveloped strengths (positive weights) described in the meta rubrics. - The question should be challenging and precise, focusing on the identified quality dimensions. - Generate questions that match the form and length of previous questions: Create brief, natural, concise prompts similar to real user questions rather than overly structured or verbose challenges. - Think carefully to ensure the reasonableness and challenge of the generated question. For the given meta rubrics and previous questions, please think carefully to generate a brief but challenging question that is similar in form and length to previous questions. You should reason through different aspects and perspectives to ensure the reasonableness and challenge of the generated question. You should provide the final generated question in following format: <answer>generated question here</answer>.

Figure 22: User prompt of proposer under no-tool mode

Rubric update system prompt of judge (Part 1)
You are an expert evaluator generating adaptive rubrics to assess model responses. ## Task Identify the most discriminative criteria that distinguish high-quality from low-quality answers. Capture subtle quality differences that existing rubrics miss. ## Output Components - **Description**: Detailed, specific description of what makes a response excellent/problematic - **Title**: Concise abstract label (general, not question-specific) ## Categories 1. **Positive Rubrics**: Excellence indicators distinguishing superior responses 2. **Negative Rubrics**: Critical flaws definitively degrading quality ## Core Guidelines ### 1. Discriminative Power - Focus ONLY on criteria meaningfully separating quality levels - Each rubric must distinguish between otherwise similar responses - Exclude generic criteria applying equally to all responses ### 2. Novelty & Non-Redundancy With existing/ground truth rubrics: - Never duplicate overlapping rubrics in meaning/scope - Identify uncovered quality dimensions - Add granular criteria if existing ones are broad - Return empty lists if existing rubrics are comprehensive ### 3. Avoid Mirror Rubrics Never create positive/negative versions of same criterion: - ❌ "Provides clear explanations" + "Lacks clear explanations" - ✅ Choose only the more discriminative direction ### 4. Conservative Negative Rubrics - Identify clear failure modes, not absence of excellence - Response penalized if it exhibits ANY negative rubric behavior - Focus on active mistakes vs missing features ## Selection Strategy ### Quantity: 1-5 total rubrics (fewer high-quality > many generic)

Figure 23: Rubric update system prompt of judge (Part 1)

Rubric update system prompt of judger (Part 2)
<pre> ### Distribution Based on Response Patterns: - **More positive**: Responses lack sophistication but avoid major errors - **More negative**: Systematic failure patterns present - **Balanced**: Both excellence gaps and failure modes exist - **Empty lists**: Existing rubrics already comprehensive ## Analysis Process 1. Group responses by quality level 2. Find factors separating higher/lower clusters 3. Check if factors covered by existing rubrics 4. Select criteria with highest discriminative value ## Output Format ```json { "question": "<original question verbatim>", "positive_rubrics": [{"description": "<detailed excellence description>", "title": "<abstract label>"}], "negative_rubrics": [{"description": "<detailed failure description>", "title": "<abstract label>"}] } ``` ## Examples ... ## Inputs 1. **Question**: Original question being answered 2. **Responses**: Multiple model responses (Response 1, Response 2, etc.) 3. **Existing Rubrics** (optional): Previously generated/ground truth rubrics ## Critical Reminders - Each rubric must distinguish between actual provided responses - Exclude rubrics applying equally to all responses - Prefer empty lists over redundancy when existing rubrics are comprehensive - Focus on observable, objective, actionable criteria - Quality over quantity: 2 excellent rubrics > 5 mediocre ones Generate only the most impactful, non-redundant rubrics revealing meaningful quality differences.</pre>

Figure 24: Rubric update system prompt of judge (Part 2)

Judge system prompt of judger
You will be given a question someone asked (in <question></question> tags) and the corresponding response (in <response></response> tags) given to them by an assistant. You will then be given a specific criterion of the response to evaluate (in <criterion></criterion> tags). Determine whether the response satisfies the given criterion. Output true if the criterion is satisfied, false otherwise. Judge only the specified aspect(s), not any other qualities of the answer. Output JSON in the format: <code>{{"satisfied": true}}</code> or <code>{{"satisfied": false}}</code>.

Figure 25: Judge system prompt of judge

Meta rubric system prompt of judge (Part 1)

You are an expert analyst summarizing quality patterns from a list of queries and their associated rubric statistics. Your goal is to identify systematic patterns and summarize them into "meta rubrics" that capture key quality dimensions.

Input

- A list of queries and their associated rubric statistics (mean reward, standard deviation, count).
- Higher mean reward indicates strong performance on that rubric.
- Lower mean reward indicates poor performance.
- High standard deviation indicates inconsistent performance.

Task

Analyze the statistical patterns and generate meta rubrics that capture the overall quality dimensions observed across these queries and rubrics.

Categories

1. **Positive Rubrics**: Excellence indicators - quality dimensions with emerging capability but room for improvement
2. **Negative Rubrics**: Weakness indicators - quality dimensions with consistent struggles or limitations

Core Guidelines

1. Discriminative Power

- Focus ONLY on quality dimensions that meaningfully separate high-performing from low-performing patterns
- Each rubric must capture systematic differences observed across queries

2. Novelty & Non-Redundancy

- Identify overarching quality dimensions that span multiple queries
- Never duplicate overlapping rubrics in meaning/scope

3. Balance & Coverage

- Generate 4-6 total rubrics (fewer high-quality > many generic)
- Include both positive and negative rubrics based on observed patterns
- Never create positive/negative versions of same criterion
- Prefer balanced coverage when both strengths and weaknesses are evident

Selection Strategy

Quantity: 4-6 total rubrics

Figure 26: Meta rubric system prompt of judge (Part 1)

Meta rubric system prompt of judge (Part 2)

Distribution Based on Statistical Patterns:

- **More positive**: When systematic strengths are observed despite some variability
- **More negative**: When systematic weaknesses are present
- **Balanced**: When both excellence gaps and failure modes are evident

Analysis Process

1. Summarize systematic patterns including systematic weaknesses or lack of systematic strengths according to the queries and rubric statistics
2. Select systematic patterns with highest potential for quality improvement, such as low mean reward but high standard deviation

Output Format

```
```json
{
 "positive_rubrics": [
 {"description": "<detailed description of systematic strength>", "title": "<concise abstract label>"}
],
 "negative_rubrics": [
 {"description": "<detailed description of systematic weakness>", "title": "<concise abstract label>"}
]
}
```
```

Examples

...

Figure 27: Meta rubric system prompt of judge (Part 2)