Large Language Model-based Human-Agent Collaboration for Complex Task Solving

Xueyang Feng[1,2*], **Zhi-Yuan Chen**[1,2*], **Yujia Qin**[3], **Yankai Lin**[1,2†]
Xu Chen[1,2†], **Zhiyuan Liu**[3], **Ji-Rong Wen**[1,2]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2] Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
[3] Department of Computer Science and Technology, Tsinghua University, Beijing, China
{xueyangfeng, zhiyuanc2001, yankailin, xu.chen}@ruc.edu.cn

Abstract

In recent developments within the research community, the integration of Large Language Models (LLMs) in creating fully autonomous agents has garnered significant interest. Despite this, LLM-based agents frequently demonstrate notable shortcomings in adjusting to dynamic environments and fully grasping human needs. In this work, we introduce the problem of LLM-based human-agent collaboration for complex task-solving, exploring their synergistic potential. To tackle the problem, we propose a **Re**inforcement Learning-based **H**uman-**A**gent **C**ollaboration method, **ReHAC**, which trains a policy model designed to determine the most opportune stages for human intervention within the task-solving process. We conduct experiments under ***real*** and ***simulated*** human-agent collaboration scenarios. Experimental results demonstrate that the synergistic efforts of humans and LLM-based agents significantly improve performance in complex tasks, primarily through well-planned, limited human intervention. Datasets and code are available at: `https://github.com/XueyangFeng/ReHAC/`.

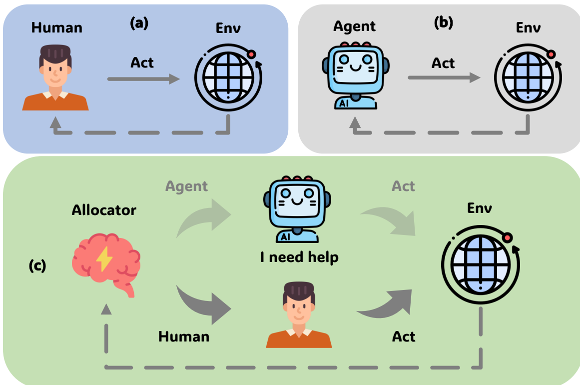

Figure 1: Different Levels of Automation. **(a) No automation:** Tasks are entirely performed by humans. **(b) Full automation:** Tasks are completely executed by agents without human intervention. **(c) Conditional automation:** Humans are required only for specific subtasks, without continuous monitoring.

1 Introduction

In today's increasingly complex world, humans are confronted with multifaceted tasks stemming from technical, social, and economic domains. Solving these complex tasks necessitates not only human interaction with the environment but also intricate decision-making processes. To alleviate human workload and enhance the automation of tasks in both professional and personal spheres, researchers have been actively developing advanced tools for human assistance (Zawacki-Richter et al., 2019; Amershi et al., 2019).

Recently, the emergence of Large Language Models (LLMs) (Touvron et al., 2023; Team et al., 2023; Brown et al., 2020; Achiam et al., 2023) has marked a significant milestone. LLMs' remarkable abilities in task understanding, planning, and reasoning (Zhao et al., 2023) have given rise to the development of LLM-based autonomous agents (Wang et al., 2023a; Yao et al., 2022; Shinn et al., 2023). These agents are designed to leverage the LLMs' capabilities to assist humans in solving complex tasks autonomously. The LLMs' capabilities enable them to effectively navigate and address the complexities encountered in real-world scenarios, thereby offering substantial support in human decision-making processes of task-solving.

Despite the remarkable progress of LLM-based agents, there remains a notable gap in their intelligence level to handle complex and dynamic real-world tasks with human-like proficiency. This limitation poses a significant challenge to their practicality in real-world applications, especially in scenarios where high accuracy is crucial, such as the legal or financial domains. Addressing this challenge extends beyond just enhancing the agents' capabilities. Incorporating human intuition and

[*] Equal Contribution. The order is determined by dice rolling.
[†] Corresponding Authors.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1336–1357
November 12-16, 2024

wisdom is equally vital for the effective management of these intricate and evolving tasks, offering a complementary approach to the limitations of current agent technologies.

In this work, we introduce the problem of **LLM-based human-agent collaboration for complex task solving**, aiming to augment the capabilities of LLM-based agents by integrating human intuition and wisdom. The idea is analogous to the evolution in autonomous driving technology, which has been categorized into varying levels of autonomy, ranging from no automation, conditional automation to full automation (Khan et al., 2022; SAE International, 2021). Referring to this framework, we define different levels of human-agent collaboration for complex task solving, as illustrated in Figure 1. Instead of aiming for full automation, human-agent collaboration under the paradigm of conditional automation enables humans to intervene the complex task-solving when necessary, while agents handle most of the sub-tasks. Applying this conditional automation mode to LLM-based agents offers a practical path for their deployment in real-world scenarios, acknowledging the current limitations in their cognitive capabilities. Some researchers have made preliminary attempts, by designing heuristic rules or specialized prompts to determine the stages at which agents should seek human assistance (Cai et al., 2023; Wu et al., 2022a; Mehta et al., 2023; Wang et al., 2023b). However, these rule-based or prompt-driven approaches are heavily reliant on specific application contexts and lack universality. They often demand a deep understanding of the domain and substantial experience from the designers, otherwise, suboptimal design choices can lead to reduced performance. Apart from that, a standardized formal framework and universally accepted paradigm for leveraging large language models (LLMs) in human-agent collaboration is still lacking.

To overcome the aforementioned challenges, we propose a **Re**inforcement Learning-based **H**uman-**A**gent **C**ollaboration method(**ReHAC**), which is a learnable general framework aimed at dynamically identify the most advantageous stages for human intervention during the task-solving process. In the experiments, we first conduct **real** human-agent experiments on the HotpotQA dataset. To further assess the efficacy of our method, we conduct **simulated** human-agent collaboration experiments on the HotpotQA, StrategyQA, and InterCode datasets. In addition, we also analyze the generalization of our ReHAC method on different prompt frameworks and collaboration paradigms. To summarize, our experimental results indicate that with a policy model learned from limited data, ReHAC can effectively allocate human intervention in human-agent collaboration scenarios, thereby achieving a balance between effectiveness and efficiency (average 25.8% relative improvement over baselines on the HotpotQA dataset).

2 Approach

In this section, we first formulate the problem of human-agent collaboration for complex task solving, and then introduce our proposed ReHAC method in detail.

2.1 Preliminary and Problem Formulation

Complex task-solving, inherently necessitating multi-step planning and reasoning, is conventionally formalized as a multi-step decision-making problem. Historically, complex task-solving was predominantly achieved through **human-driven methods**. These methods leveraged human cognitive capabilities to determine the suitable action in each step. Formally, considering a complex task q, it is traditionally solved via a sequence of actions $(a_1, a_2, \cdots a_n)$, with each action determined by human decision-making, expressed as:

$$a_t = \text{Human}(q, s_t), \tag{1}$$

where $s_t = (a_1, o_1, \cdots, a_{t-1}, o_{t-1})$ denotes the history information of task state at step t and o_t is the observation after a_{t-1} is proceeded.

The advent of LLMs has brought a paradigm shift in this arena. Their impressive understanding and reasoning abilities have prompted research into LLM-based agents for complex task-solving, thereby enhancing the level of automation in task-solving. These **agent-driven methods** (e.g., ReAct (Yao et al., 2022)), leverage LLM-based agents to supplant human decision-making. This shift is represented as:

$$a_t = \text{Agent}(q, s_t). \tag{2}$$

This evolution of such AI-driven techniques provides a way to the automation of complex task-solving.

However, limited by the current intelligence level of LLMs, full automation based on agent-driven methods is not yet feasible in practical scenarios (Kiseleva et al., 2022; Mehta et al., 2023).

Inspired by autonomous driving (Cui et al., 2024; Fu et al., 2024; Bastola et al., 2024), we propose the problem of **LLM-based human-agent collaboration for complex task solving** and explore the dynamics and efficacy of the **human-agent collaborative methods** for complex task solving. We first explore a specific form of human-agent collaboration: humans intervene in the complex task-solving process when necessary. Formally, we need to determine whether a human or an agent makes decisions based on the actions' complexity and contextual changes, i.e.,

$$a_t = \text{Human}(q, s_t) \quad \text{or} \quad \text{Agent}(q, s_t), \quad (3)$$

It is generally perceived that direct human intervention in decision-making, particularly in real-world scenarios, incurs higher costs and diminishes the system's automation level (Cai et al., 2023; Wang et al., 2023b). On the other hand, human intervention plays an important role in enhancing task performance and flexibility. Therefore, the objective of human-agent collaboration is to enhance the effectiveness of complex task-solving with minimal reliance on human decision-making. One key challenge is to **determine the stages in the task-solving process where human intervention is most beneficial and effective, aligning with the goal of minimizing human involvement while maximizing task performance**.

2.2 ReHAC

In this work, we propose a Reinforcement learning-based Human-Agent Collaboration method, ReHAC. It formulates the human-agent collaboration problem as a Markov Decision Process (MDP) framework, represented by the tuple $(S, \mathcal{A}, P, R, \gamma)$, where S is the set of states, $\mathcal{A}$ is the set of actions, $P : S \times \mathcal{A} \times S$ is the state transition probabilities, R serves as the reward function, and γ the discount factor.

For each action $a_t \in \mathcal{A}$, we define it as a tuple $(a_t^{collab}, a_t^{task})$, where $a_t^{collab} \in \{0, 1\}$ indicates the subtask is allocated to an agent or a human, and a_t^{task} is the task action determined by agent or human:

$$a_t^{collab} \sim \pi_{\theta_1}^{collab}(a_t^{collab}|s_t)$$

$$a_t^{task} \sim \begin{cases} \pi_{\theta_2}^{task}(a_t^{task}|s_t), & \text{if } a_t^{collab} = 0; \\ \pi_{\text{Human}}^{task}(a_t^{task}|s_t), & \text{otherwise}, \end{cases} \quad (4)$$

where $\pi_{\theta_1}^{collab}$ is the collaboration policy model, $\pi_{\theta_2}^{task}$ is the agent-based task policy model, and $\pi_{\text{Human}}^{task}$ is the human task policy[1].

To balance the maximization of task performance and the cost of human intervention, we define the reward function as:

$$R(s, a) = T(s, a) - \lambda C(s, a), \quad (5)$$

where $T(s, a)$ is the measure of expected task rewards received after taking action a in state s, $C(s, a)$ is the number of human interventions in the trajectory after taking action a, λ is a hyper-parameter that serves as a penalty coefficient of the number of human interventions. We utilize Monte-Carlo estimation to compute this reward function.

Optimization: Following the policy gradient algorithm (Schulman et al., 2017), we optimize the advantage function:

$$\mathcal{J}(\pi_\theta) = \mathbb{E}[\frac{\pi_\theta(a|s)}{\pi_{\text{beh}}(a|s)} A(s, a)],$$
$$A(s, a) = R(s, a) - \frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} R(s, a'), \quad (6)$$

which aims to find an optimal policy π_θ that ensures the maximization of task rewards while minimizing the human intervention costs, and $\theta = [\theta_1, \theta_2]$. In order to increase sample efficiency and reduce costs, we propose to convert policy gradient to offline form for training. The behavioral policy π_{beh} represents the policy for collecting offline training data.

The gradient of our objective function is:

$$\nabla_\theta \mathcal{J}(\pi_\theta) = \sum_{(s,a)} w(s, a) \nabla_\theta \log \pi_\theta(a|s) A(s, a),$$
$$w(s, a) = \text{Clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\text{beh}}(a|s)}, 1 - \epsilon, 1 + \epsilon\right), \quad (7)$$

where the clip function limits the importance sampling term to the interval $1 - \epsilon$ to $1 + \epsilon$[2]. Moreover, we incorporate an entropy regularization term $H(\cdot)$. This term encourages the policy to explore a variety of actions, thereby preventing the policy from becoming too deterministic and overfitting to the training data. Finally, the gradient of objective

[1]The agent-based task policy model can be a closed-source model like ChatGPT or an open-source model.

[2]The clip function is calculated by $w(s, a) = \max(1 - \epsilon, \min(\frac{\pi_\theta(a|s)}{\pi_{\text{beh}}(a|s)}, 1 + \epsilon))$.

function is as follows:

$$\nabla_\theta \tilde{\mathcal{J}}(\pi_\theta) = \nabla_\theta \mathcal{J}(\pi_\theta) + \alpha \nabla_\theta H(\pi_\theta(\cdot|s)),$$
$$H(\pi_\theta(\cdot|s)) = -\sum_a \pi_\theta(a|s) \log \pi_\theta(a|s). \quad (8)$$

3 Experiments

3.1 Experimental Setup

Datasets Following Yao et al. (2022); Shinn et al. (2023); Liu et al. (2023); Xu et al. (2023), we evaluate the efficacy of our method on two question answering datasets: HotpotQA and StrategyQA, and a coding dataset: InterCode. We provide more details about the datasets in the Appendix A.1.

Prompt Framework The core of our method lies in incorporating humans to enhance the performance of LLM-based agents. Since this method is independent of prompt framework design, it remains adaptable across various prompt frameworks. In our experiments, we set humans and agents to solve tasks under the ReAct framework (Yao et al., 2022) for question-answering datasets and "Try Again" framework (Yang et al., 2023) for the InterCode dataset. These two frameworks represent two fundamental paradigms in agent prompting. ReAct incorporates multi-step reasoning and tool utilization, while "Try Again" builds upon ReAct by guiding actions based on external rewards. We also conduct experiments under more complex prompt framework Reflexion (Shinn et al., 2023) in Section 3.3. More details about the prompt framework can be found in the Appendix A.3.

Implementation details We use Llama2 (Touvron et al., 2023) as the collaboration policy model $\pi_{\theta_1}^{collab}$ and use Low-Rank Adaptation (LoRA, Hu et al. (2021)) methods to train the policy model. In all experiments, we utilized ChatGPT (gpt-3.5-turbo-0613) to simulate the agent policy $\pi_{\theta_2}^{task}$. More model implementation and data collection details can be found in Appendix A.2.

Reward Calculation For all datasets, the final reward is computed as equation (5). For question answering datasets, we choose the F1 score as the task reward $T(s, a)$. For the InterCode dataset, following Yang et al. (2023), we use Intersection over Union as the task reward $T(s, a)$.

Baselines Our method focuses on better coordinating human-agent collaboration within the same agent framework, making it orthogonal to other agent frameworks and not suitable for direct comparison. Therefore, we designed several baseline methods for our experiments: 1) Agent-only which carries out all actions by agents. 2) Human-only, which conducts all actions by humans. 3) Random, which selects an agent or human randomly at a probability of 50% to perform each action. 4) Prompt, which prompts the agent to actively decide whether the action is executed by itself or a human. 5) Imitation Learning (IL), which trains the policy model to decide whether the action should be finished by an agent or human by the IL method. More details about baselines can be found in the Appendix A.4.

3.2 Overall Results

In this section, we first employ real users and conduct extensive experiments on the HotpotQA dataset to verify the effectiveness of our proposed ReHAC method. In addition, to further verify the effectiveness of our method from multiple angles, we constructed a simulated user environment to conduct extended experiments.

Real Human-Agent Experiments Figure 2(a) shows the evaluation results of human-agent collaboration on the HotpotQA dataset. From the figure, we can observe that all human-agent collaboration methods outperform Human-only and Agent-only methods. This underscores the importance of collaborating human and agent in complex task-solving for getting higher reward. In addition, ReHAC$_{\text{Human}}$ achieves the best performance compared with prompt-based and random-based method in achieving higher rewards. Specifically, when $\lambda = 0.06$, ReHAC achieves a higher reward with approximately 30% more human interventions compared with the prompt-based baseline; when $\lambda = 0.1$, it also achieves a reward improvement with about 20% less human interventions. This indicates that our ReHAC method can dynamically introduce human intervention in real human-agent collaboration scenarios, thereby achieving a balance between effectiveness and efficiency.

Focusing on ReHAC$_{\text{Human}}$, we observe that as λ increases, the human intervention rate[3] (HIR) of ReHAC$_{\text{Human}}$ gradually decreases. This trend suggests that a higher human penalty coefficient elevates our policy model's "threshold" for assigning actions to humans. Simultaneously, the decrease of

[3]The formula for calculating the human intervention rate is in Appendix A.5.

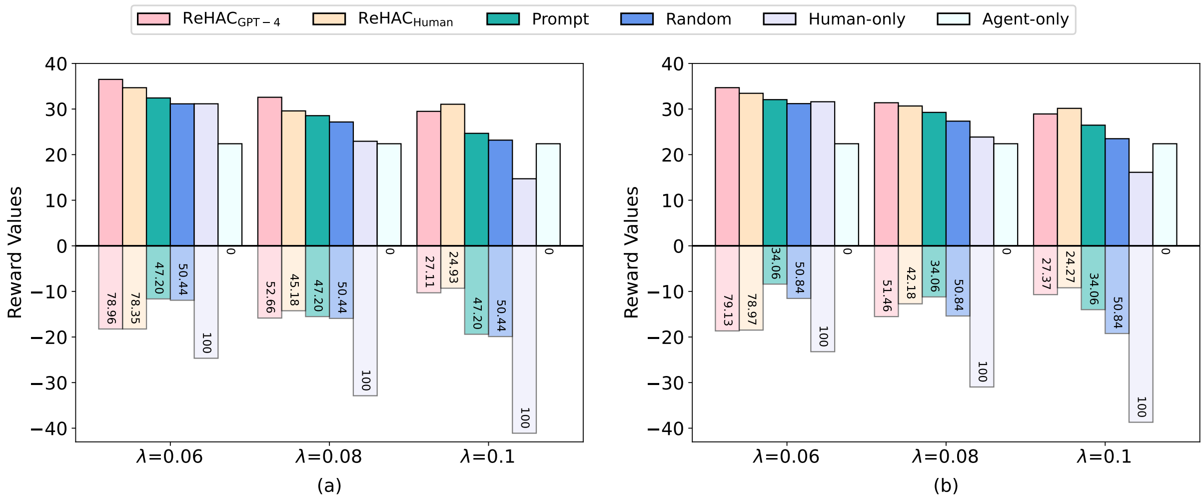

Figure 2: (a) Human-agent collaboration evaluation. (b) GPT-4-agent collaboration evaluation. The bars above the 0-axis represent the reward R, the bars below the 0-axis represent the human intervention cost λC, and the entire columns, composed of the bars above and below the 0-axis, represent the task reward T. Numbers within the bars means the human intervention rate (%). ReHAC$_{\text{GPT-4}}$ and ReHAC$_{\text{Human}}$ represent the policy model trained on GPT-4-agent and human-agent collaboration datasets, respectively. ReHAC outperforms other baselines in human-agent collaboration scenarios.

the HIR correspondingly results in a deterioration of human-agent interaction performance.

Simulated Human-Agent Experiments Due to the high cost of hiring annotators to label real human-agent collaboration data, it is costly for us to collect human-agent collaboration data on more datasets and, as a result, validate the efficacy of our method in broader scenarios. We instead use GPT-4 (gpt-4-0613) to build a simulation environment and make it collaborate with agents to solve tasks. This setup enables us to collect more "human-agent" collaboration data at a reasonable cost.

To verify the feasibility of using GPT-4 to simulate humans to collect "human-agent" collaboration data, we learn ReHAC on the HotpotQA GPT-4-agent collaboration data, named as ReHAC$_{\text{GPT-4}}$ and test its performance in the real human-agent collaboration environment. From Figure 2(a), we can see that ReHAC$_{\text{GPT-4}}$ exhibits better performance compared to ReHAC$_{\text{Human}}$ in human-agent collaboration when $\lambda = 0.06$ and 0.08. We suppose that this is possibly attributed to individual differences among humans, leading to a distribution variance in the human-agent collaboration data, while GPT-4-agent collaboration data exhibits higher consistency and lower variance. This makes ReHAC$_{\text{GPT-4}}$ learn the collaboration signal more easily, and thus is more stable and performs better.

To further reduce costs and observe the reward variation of ReHAC during the training process, we use GPT-4 to simulate humans in the evaluation phase. Figure 2(b) shows the evaluation results when using GPT-4 to simulate humans for collaboration. Comparing the results in Figure 2(a) and (b), we notice that the relative performance of various methods is generally consistent in both human-agent collaboration and GPT-4-agent collaboration. For example, the rewards R of ReHAC consistently surpass those of the Prompt method, and both ReHAC and the Prompt method outperform the Random method. This demonstrates the viability of using GPT-4 to simulate humans for evaluation.

Learning Curves Figure 3 shows the learning curves during the training process. The curves are obtained by assessing the policy model's rewards on the trainset and testset every 5 steps. From the figure, we can observe that (1) the rewards of ReHAC gradually increase during the training process, indicating that ReHAC can progressively identify suitable points to introduce human interventions. (2) While the IL method achieves high rewards on the trainset, it performs poorly on the testset. In contrast, for ReHAC, the rewards on both the trainset and the testset continuously increase as training progresses. This suggests our RL-based learning method learns a more generalized human-agent collaboration strategy compared to directly learn-

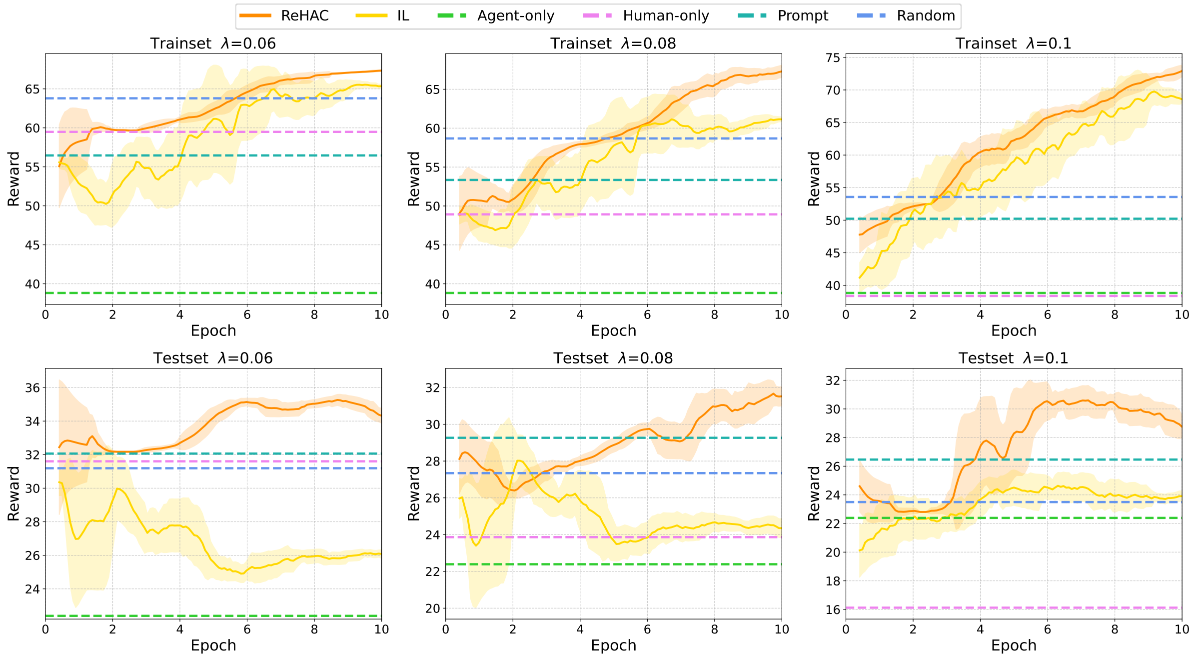

Figure 3: Reward R variations of different methods during the training process on HotpotQA dataset. Here we set the human intervention penalty coefficient λ to 0.06, 0.08, and 0.1. Curves of ReHAC and IL are averaged over 15 points, with shadows indicating the variance.

Figure 4: Reward R variations during the training process on StrategyQA and InterCode datasets. Curves of ReHAC and IL are averaged over 15 points, with shadows indicating the variance.

ing the optimal strategy with the imitation learning method.

3.3 Performance on Different Dataset

In this part, we train and test ReHAC method on StrategyQA and InterCode datasets in the GPT-4 simulation environment. For all experiments, we fix the parameter $\lambda = 0.08$. Throughout the training phase, we evaluate the policy model's rewards on the trainset and testset every 5 steps. Experimental results are shown in Figure 4. From the figure, we observe that: (1) Our proposed ReHAC method achieves higher reward scores compared to other baselines on all datasets. This validates the effectiveness of our approach across a broader range of datasets. (2) Both ReHAC and IL exhibit low variance and stability during the training process. Although our method and the IL method show a continuous reward increase during the training process, ReHAC can ultimately achieve higher rewards compared to the IL method. This indicates that our reinforcement learning-based method can provide more valuable guidance to the policy model $\pi_{\theta_1}^{collab}$, enabling it to determine when to introduce human interventions and consequently achieving higher rewards.

In summary, our method demonstrates superior performance across all datasets, affirming its ability to achieve an optimal balance between efficiency and effectiveness.

3.4 Generalization Analysis

In order to verify the adaptability and generalization of our method, we conducted two dimensions of generalization experiments: generalization research on the prompt framework and generalization

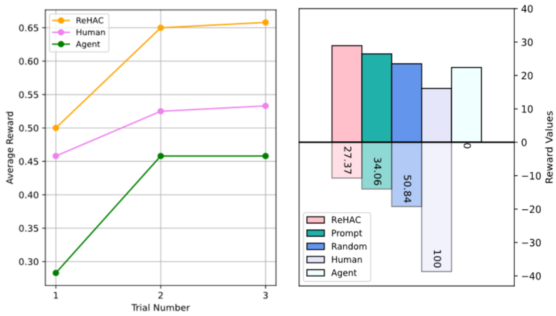

Figure 5: **Left**: The average reward for each trial in the Reflexion framework. **Right**: Evaluation results where humans are required to modify the actions of agents.

Model	HIR (%)	Task Reward T	Reward R
HotpotQA			
Llama2-7B	51.46	46.90	31.38
Llama2-13B	47.64	46.78	32.22
InterCode			
Llama2-7B	4.15	62.00	60.08
Llama2-13B	3.10	60.00	58.56

Table 1: Experimental results regarding different model scales.

research on collaboration paradigms.

Prompt Framework Generalization To futher evaluate if human intervention would be effective within more complex prompting frameworks, we conduct experiments using the Reflexion (Shinn et al., 2023) framework. We directly employ the model checkpoint trained under the ReAct framework to determine whether each step should be completed by a human or an agent in each trial. The experiment is implemented on the HotpotQA dataset with $\lambda = 0.1$, and results are shown in Fig 5. As shown in the figure, our ReHAC method outperforms both Agent-Only and Human-Only across all trials. The experimental results demonstrate the ability of ReHAC to generalize across different prompt frameworks.

Collaboration Paradigm Generalization Previous experiments have mainly focused on humans directly replacing agents in action. In order to prove that our method can be generalized to other collaboration paradigms, we construct a new form of collaboration where the tasks completed by agents are handed over to humans for modification. For each action $a_t \in \mathcal{A}$, we define it as a tuple $(a_t^{collab}, a_t^{task})$, where $a_t^{collab} \in \{0, 1\}$ indicates whether the subtask completed by the agent should be assigned to a human to modify, and a_t^{task}

Method	HIR (%)	Task Reward T	Reward R
HotpotQA			
Agent-only	0.0	22.39	22.39
Human-only	100.0	54.82	23.86
ReHAC (single-task)	51.46	46.90	31.38
ReHAC (multi-task)	32.65	41.06	30.82
StrategyQA			
Agent-only	0.0	60.00	60.00
Human-only	100.0	68.00	43.36
ReHAC (single-task)	20.47	66.00	61.12
ReHAC (multi-task)	14.14	65.00	61.64

Table 2: Experimental results on multi-task training.

represents the action performed by the agent, with or without human modification:

$$
\begin{aligned}
a_t^{agent} &\sim \pi_{\theta_2}^{task}(s_t) \\
a_t^{collab} &\sim \pi_{\theta_1}^{collab}(a_t^{collab} | a_t^{agent}, s_t) \\
a_t^{task} &\sim \begin{cases} a_t^{agent}, & \text{if } a_t^{collab} = 0 \\ \pi_{\text{Human}}^{modify}(a_t^{task} | a_t^{agent}, s_t), & \text{otherwise} \end{cases}
\end{aligned}
\tag{9}
$$

where $\pi_{\theta_1}^{collab}$ is the collaboration policy model, $\pi_{\theta_2}^{task}$ is the agent-based task policy model, and $\pi_{\text{Human}}^{modify}$ is the human modify policy. Experimental results are shown in Fig 5. As evidenced by the experimental results, our ReHAC method achieves higher rewards compared to other baselines. This suggests that our learned collaboration policy model has successfully learned to introduce human modifications at the opportune stage. Furthermore, the experimental results demonstrate the adaptability of our method to different collaboration paradigms.

3.5 Scaling Analysis of Policy Model

In this section, we analyze the impact of the model scale on the performance of the policy model. Here, we set $\lambda = 0.08$ and conduct experiments on HotpotQA and InterCode datasets. As shown in Table 1, the Llama2-7B model performs competitively with the Llama2-13B model. This suggests that the 7B model is already proficient in handling the human-agent collaboration task, and the benefit of increasing the size of the model is slight. We will explore smaller policy model size in the future.

3.6 Multi-Task Learning

To demonstrate the effectiveness of our method across different tasks, we conduct multi-task training on the HotpotQA and StrategyQA datasets. In the experiment, we set $\lambda = 0.08$. The results are

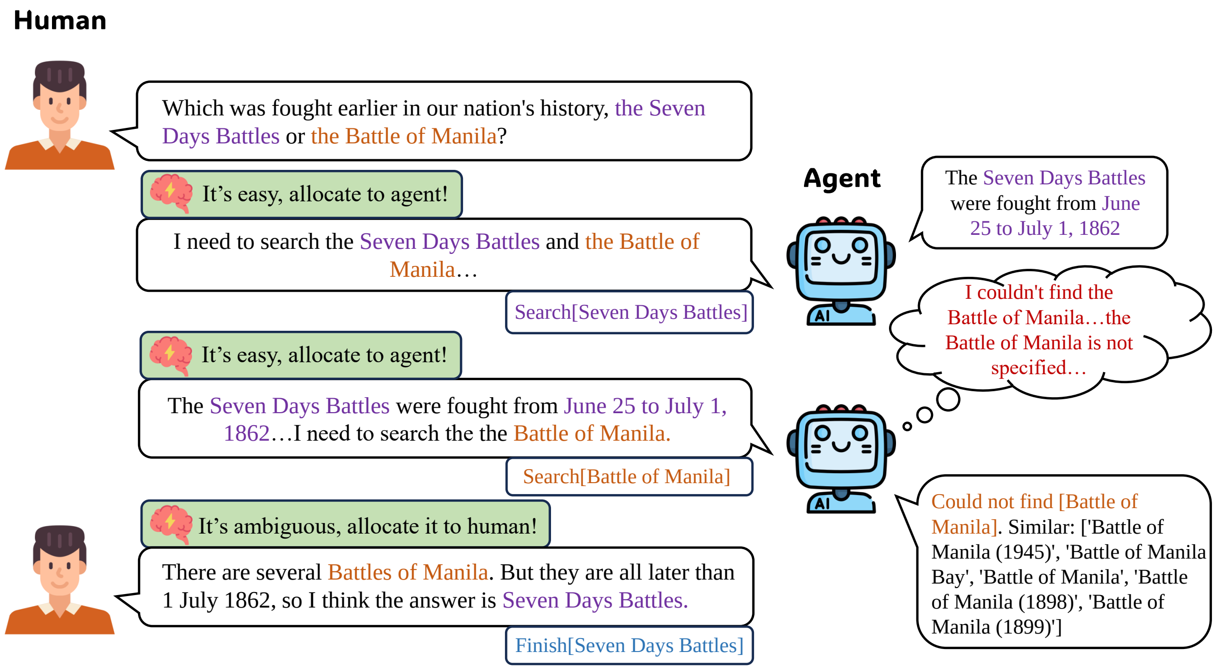

Figure 6: Case Study. When the agent completes the task, the third step cannot be answered due to the ambiguity of the problem identified; using our method, the first two simple retrieval tasks are assigned to the agent to complete, while the third step is assigned to humans. Humans can complete the correct answer through bold speculation

illustrated in Table 2. As shown in the table, our ReHAC method achieves similar rewards to the single-task scenarios under the multi-task scenarios (for example, on the StrategyQA dataset, the rewards of ReHAC in multi-task and single-task scenarios are 61.64 and 61.12, respectively), and it outperforms both the Agent-only and Human-only baselines (the rewards of Agent-only and Human-only on StrategyQA are 60.00 and 43.46, respectively). This demonstrates that our method is still effective in multi-task scenarios.

3.7 Case Study

In this section, we analyze a large amount of ReHAC cases from real user feedback. we provide cases where ReHAC can solve three dilemmas of LLM-agent: **missing information, task ambiguity, and dead loop**. Here, we give a specific case about information ambiguity for detailed analysis. More cases for each type of situation are in the Appendix A.6. As illustrated in Figure 6, the task is to determine which historical event, the Seven Days Battles or the Battle of Manila, occurred first. When given the entire problem, the agent accurately determines the date of the Seven Days Battles but encounters multiple entries for the Battle of Manila, resulting in ambiguity. Consequently, the agent deems the query ambiguous and opts to respond with "unknown". On the contrary, our ReHAC method requires the human intervention in this situation. Upon examining the related entries, the human observes that all mentioned dates for the Battle of Manila occurs after to July 1, 1862. Based on this insight, he conjectures that the Seven Days Battles occurred first. Although this conjecture is not absolutely certain, it represents the most likely decision based on the available information. Thus, our ReHAC method returns a correct response "Seven Days Battles". This case also highlights an insightful aspect of our research into LLM-based agents: Researchers are committed to eliminating hallucinations in large language models (LLMs) to create rigorous and accurate intelligent agents. However, many tasks require imagination and intuition, making it crucial to integrate human creative thinking through human-agent collaboration at this juncture.

4 Discussion

In this section, we propose two extended research directions to enhance the effectiveness, safety, and intelligence of human-agent collaboration:

Development Stages of LLM-based Agents Inspired by the L1 to L5 grading model in autonomous driving, we suggest adapting this framework for LLM-based human-agent collaboration.

It offers a clear structure to assess the current development stage of human-agent technologies and guide future research. While LLM agents have not reached high or full automation, this framework is crucial for identifying key technologies and challenges. However, our research indicates a significant gap before LLM agents achieve full automation (L5). Effective human-agent collaboration could be a bridge towards this goal.

Safety and Super Alignment Safety is paramount in human-agent collaboration, particularly in high-risk scenarios. It's vital to explore methods to secure the collaboration process and mitigate risks. Moreover, with the potential of LLM-based agents evolving into superintelligence, effective collaboration becomes increasingly crucial. This collaboration is key, as it not only allows humans to guide ethical and safety decisions but also ensures the alignment of LLM-based agents' objectives with human interests.

5 Related Work

LLM-based Agent LLM-based agents (Yao et al., 2022; Shinn et al., 2023; Qin et al., 2023; Wang et al., 2023a), which can interact with the environment and select subsequent actions based on environment feedback, have been applied in many domains, including web navigation (Nakano et al., 2021; Cheng et al., 2024; He et al., 2024), software engineering (Qian et al., 2023; Hong et al., 2023), and robotics (Wang et al., 2024; Mahadevan et al., 2024). However, current LLM-based agents still perform poorly on some complex tasks. This work aims to introduce human interventions and enable humans and agents to collaboratively address complex tasks, thereby achieving improved task performance.

Human-Agent Collaboration Human-Agent Collaboration (HAC) involves improving human interactions with AI systems and robots (Wang et al., 2021; Wu et al., 2022b). Recent advancements emphasize the importance of human feedback in enhancing the capabilities of language model-based agents. Studies have developed heuristic rules(Cai et al., 2023; Wu et al., 2022a; Mehta et al., 2023) and specialized prompts (Huang et al., 2022; Wang et al., 2023b) to encourage these agents to seek human input, fostering a more collaborative dynamic. Effective design of these elements, crucial for handling complex tasks, relies on the designer's expertise. Our research aims to create a generalized, learnable method to coordinate human and AI collaboration through direct planning.

6 Conclusion

In this paper, we explore the integration of human expertise and the computational power of large language models (LLMs) in complex decision-making tasks. We introduce a reinforcement learning approach, ReHAC, for human-agent collaboration. ReHAC uses a learnable policy to identify key points for human intervention in task resolution. Experimental results suggest that ReHAC outperforms traditional heuristic and prompt-based methods in human-agent tasks, providing a viable framework for applying LLMs in real-world scenarios.

Ethical Considerations and Limitations

The objective of this work focuses on human-agent collaboration, which requires humans to interact with LLM-based agents. We acknowledge that agents are likely to output some hallucinations and misleading information, and it is unclear how these contents impact humans. Additionally, all datasets used in this work are publicly available, and therefore, there are no data privacy concerns. All data collected will be used for research purposes only.

The limitations of this paper can be summarised in two aspects:

1) Our research primarily focuses on the use of 7B and 13B scale models as policy models for task allocation. Future work will investigate the feasibility of smaller models in carrying out these tasks, aiming to maintain performance while reducing resource consumption.

2) This study is based on the assumption that human performance supersedes that of agents. However, as technology advances, agents might surpass human capabilities. Future research will thus shift towards exploring human-agent collaboration models in this new context. Emphasis will be placed on assessing how human-agent collaboration can ensure the safety of agent decisions while aligning with human preferences.

Acknowledgement

We thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 62376273).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13.

Ashish Bastola, Julian Brinkley, Hao Wang, and Abolfazl Razi. 2024. Driving towards inclusion: Revisiting in-vehicle interaction in autonomous vehicles. *arXiv preprint arXiv:2401.14571*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zefan Cai, Baobao Chang, and Wenjuan Han. 2023. Human-in-the-loop through chain-of-thought. *arXiv preprint arXiv:2306.07932*.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. 2024. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 910–919.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

M. A. Khan et al. 2022. Level-5 autonomous driving—are we there yet? a review of research literature. *ACM Computing Surveys*, 55(2):Article 27.

Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, and Ahmed Awadallah. 2022. Iglu 2022: Interactive grounded language understanding in a collaborative environment at neurips 2022.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv: 2308.03688*.

Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative expressive robot behaviors using large language models. *arXiv preprint arXiv:2401.14673*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Nikhil Mehta, Milagro Teruel, Patricio Figueroa Sanz, Xin Deng, Ahmed Hassan Awadallah, and Julia Kiseleva. 2023. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. *arXiv preprint arXiv:2304.10750*.

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

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

SAE International. 2021. Sae levels of driving automation refined for clarity and international audience. https://www.sae.org/news/2021/05/sae-j3016-driving-automation-levels.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chao Wang, Stephan Hasler, Daniel Tanneberg, Felix Ocker, Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, and Michael Gienger. 2024. Large language models for multi-modal human-robot interaction. *arXiv preprint arXiv:2401.15174*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023a. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023b. Mint: Evaluating llms in multi-turn interaction with tools and language feedback.

Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. *arXiv preprint arXiv:2103.04044*.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022a. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–22.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022b. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. 2023. Lemur: Harmonizing natural language and code for language agents. *arXiv preprint arXiv:2310.06830*.

John Yang, Akshara Prabhakar, Karthik R Narasimhan, and Shunyu Yao. 2023. Intercode: Standardizing and benchmarking interactive coding with execution feedback. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Olaf Zawacki-Richter, Victoria I Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):1–27.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A Appendix

A.1 Experimental Datasets

Following Yao et al. (2022); Shinn et al. (2023); Liu et al. (2023); Xu et al. (2023), we evaluate the efficacy of our method on question answering dataset and coding datasets. (1) HotpotQA (Yang et al., 2018) is a Wikipedia-based question answering benchmark which needs model to perform multi-hop reasoning over complex questions. (2) StrategyQA (Geva et al., 2021) is a question answering benchmark with questions that need implicit reasoning. (3) InterCode (Yang et al., 2023) is an interactive coding dataset that enables agents to receive feedback from the code interpreter. In this work, we use InterCode-SQL part, which requires models to write SQL statements to fulfil the query.

A.2 Experimental Details

Model Implementation In our most experiments, we use Llama2-7b-hf[4] downloaded from Huggingface as our policy model $\pi_{\theta_1}^{collab}$. We also conduct experiments based on Llama2-13b-hf[5] model (see Section 3.3). We implement LoRA based on PEFT (Mangrulkar et al. (2022)) and set $r_{\text{LoRA}} = 16$ and $\alpha_{\text{LoRA}} = 16$ for all experiments. Based on Yao et al. (2022) and Yang et al. (2023), we set the step threshold for HotpotQA, StrategyQA, and InterCode to 7, 5, and 8, respectively. All experiments are conducted on NVIDIA A100 GPUs with 40GB memory.

Human-Agent Dataset For a real human-agent collaboration dataset, we employ a uniform sampling method where each action a_t has a 50% probability of being assigned to either a human annotator or the ChatGPT. For each question, we sample as many interaction trajectories as possible. Specifically, for each time t, we aim to sample trajectories including $a_t^{collab} = 0$ and $a_t^{collab} = 1$. Considering the diversity of responses from different annotators, we permit repeated sampling of the same trajectory during uniform sampling, which means all a_t^{collab} of two trajectories are the same. To enhance the quality of annotation, annotators are allowed to reference GPT-4's answers. We recruit 14 annotators through social media, all of whom are graduate students with strong language and reasoning skills. They are asked to annotate a total of about 2000 trajectories in four days and they get paid about \$10 an hour. They were explicitly told that the data would be used to train the model and made public and that all the labeled data was unrelated to any individual's privacy. To facilitate the annotation process, we develop a graphical user interface (GUI)[6] and provide one hour of training to annotators. The collected data details are in Table 3.

GPT-4-Agent Dataset For the dataset constructed using GPT-4 to simulate human annotation, we adopt the same sampling method as human-agent dataset collection. However, due to the uniform or near-uniform distribution of GPT-4's responses, we skip duplicate paths during uniform sampling. Collected data details are listed in Table 3.

A.3 Prompt Framework Details

In ReAct, the action space of a^{task} is {Search[entity], Lookup[keyword], and Finish[answer]}. All actions are supported by a Wikipedia web API, following the original ReAct implementation. In "Try Again" framework, agents and humans interact with the code interpreter through the action a_t and receive execution outputs from the code interpreter as observations o_t. The task-solving process ends if any one of the following conditions is satisfied: 1) the Finish[answer] action is executed actively by $\pi_{\theta_2}^{task}$ for the question answering dataset. 2) the task reward $T(s, a) = 1$ for InterCode dataset. 3) the number of actions t exceeds a pre-defined step threshold.

A.4 Baselines Details

Random We randomly choose a human or an agent to conduct action a_t at a probability of 50%.

Prompt We prompt an agent to actively decide action a_t should be finished by itself or a human. The related prompts are shown in Table 6 and Table 7. Experimental results of Random and Prompt are averaged over three repeated experiments.

Imitation Learning We select the top 50% of actions that receive the highest rewards in each state s_t as expert demonstrations. These expert demonstrations (state-action pairs) are then used to supervise the fine-tuning of the policy model. This approach allows the policy model to learn how to

[4]https://huggingface.co/meta-llama/Llama-2-7b-hf

[5]https://huggingface.co/meta-llama/Llama-2-13b-hf

[6]The GUI is as shown in Figure 7.

make decisions that get a higher return in a given state.

Dataset	Trainset		Testset
	Questions	Trajectories	Questions
HotpotQA(real)	141	1937	100
HotpotQA(sim)	141	2135	100
StrategyQA(sim)	250	2420	100
InterCode(sim)	100	2071	100

Table 3: Collected dataset details. Questions mean the number of questions we used for human-agent collaboration task. Trajectories mean the overall trajectory number we collected. (real) refers to the real human-agent collaboration dataset, and (sim) refers to the human-agent collaboration dataset collected by using GPT-4 to simulate humans.

A.5 Human Intervention Rate

We denote the number of steps completed by humans and agents in the dataset by num_h and num_a, respectively. The Human Intervention Ratio (HIR) is calculated as

$$\text{HIR} = \frac{num_h}{num_h + num_a}.$$

HIR measures the rate of human intervention. Generally, a higher HIR indicates better task performance, but it also tends to increase costs.

A.6 Case Study

In this appendix, we provide detailed case studies that further illustrate the effectiveness of the ReHAC method in addressing the challenges faced by large language model (LLM) agents in real-world applications. Each case study has been selected to demonstrate the resolution of specific dilemmas such as missing information and dead loops. These examples highlight the practical implications of our approach and offer insights into how human-agent collaboration can be optimized to enhance decision-making processes.

Case1: Missing Information. In this case (Table 8), the agent trys to determine under which U.S. President a certain American admiral served, who had collaborated with author David Chanoff and had been appointed as an ambassador to the United Kingdom. Despite employing multiple search strategies, including direct inquiries about David Chanoff's collaborators and specific searches regarding the U.S. Navy admiral's roles, the agent failed to locate any relevant information about the admiral. This lack of available data made it impossible to identify the associated President, resulting in the termination of the search effort in failure.

In the process of ReHAC task solving (Table 9), when no direct information can be searched, the next subtask is assigned to a real user. Real user added factual information: "William J. Crowe is the U.S. Navy admiral who collaborated with David Chanoff" . This critical addition of information enables agents to successfully answer queries, demonstrating how human-assisted information retrieval can bridge gaps that automated systems alone may not address. The resolution of this case not only restored the correct information but also highlighted the effectiveness of the ReHAC approach in practical applications, ultimately leading to beneficial results.

Case2: Dead Loop. The problem posed was: "Since what year has the central figure been used in the corporate branding of Singapore Airlines?" In this case (Table 10), The agent repeatedly searched for the central image of Singapore Airlines without utilizing the key information "The airline is notable for highlighting the Singapore Girl as its central figure in the corporate branding segment.". This oversight led to a redundant loop of ineffective searches. Through manual intervention (Table 11), a perceptive user capitalized on previous search attempts, accurately deduced that the "Singapore Girl" might be the central figure, and successfully used this information to resolve the query.

Case3: Dead Loop. While solving the question "Vice and Virtue" by Las Vegas band Panic!, released exactly four days after Vice and Virtue! at the Disco, is the second studio album by which Canadian rock band?" , the agent initially had difficulty in retrieving information about the album "Vices and Virtues" related to the Canadian rock band because the use of general keywords resulted in a lack of effective information and caused the search process to fall into an infinite loop (Table 12). This cycle persists because the keywords used are too broad, and searches for nouns like "Vices and Virtues" are always found instead of music albums.

ReHAC assigned this task to a real user (Table 13), and the user limited the keyword "vices and virtues (album)" to find relevant clues and break out of the wrong cycle of being unable to find information.

Experiment	α	ϵ	Learning Rate	Batch Size
HotpotQA$_{\lambda=0.06}$(GPT-4-agent, 7b)	0	0.3	3e-5	64
HotpotQA$_{\lambda=0.08}$(GPT-4-agent, 7b)	0		3e-5	
HotpotQA$_{\lambda=0.10}$(GPT-4-agent, 7b)	0		5e-5	
HotpotQA$_{\lambda=0.08}$(GPT-4-agent, 13b)	0.1		3e-5	
HotpotQA$_{\lambda=0.06}$(human-agent, 7b)	0.05		5e-5	
HotpotQA$_{\lambda=0.08}$(human-agent, 7b)	0.1		5e-5	
HotpotQA$_{\lambda=0.1}$(human-agent, 7b)	0.0		5e-5	
StrategyQA$_{\lambda=0.08}$(GPT-4-agent, 7b)	0.1		1e-5	
InterCode$_{\lambda=0.08}$(GPT-4-agent, 7b)	0		5e-5	
InterCode$_{\lambda=0.08}$(GPT-4-agent, 13b)	0.05		5e-5	

Table 4: Hyper-parameter settings for all experiments.

Methods	HotpotQA			StrategyQA			InterCode		
	HIR (%)	Task Reward	Reward	HIR (%)	Task Reward	Reward	HIR (%)	Task Reward	Reward
Agent-only	0.0	22.39	22.39	0.0	60.00	60.00	0.0	53.00	53.00
Human-only	100.0	54.82	23.86	100.0	68.00	43.36	100.0	73.00	33.72
Random	50.84	42.73	27.34	49.50	65.67	53.8	50.09	66.00	44.21
Prompt	34.06	40.46	29.26	9.14	61.33	59.12	9.94	59.33	54.69
IL	22.08	31.50	24.70	4.76	59.00	57.88	1.01	54.00	53.52
Ours	51.46	46.90	**31.38**	20.47	66.00	**61.12**	4.15	62.00	**60.08**

Table 5: ReHAC$_{\text{GPT-4}}$ Human intervention rate (HIR), task reward T, and reward R of different methods on GPT-4-agent testsets.

Imagine you are a clever planner. Given an unfinished trajectory with several steps, your task is to decide whether the next step should be carried out by ChatGPT or a human. This decision should be based on a thoughtful evaluation of the difficulty of the next step and the progress made in the current trajectory. Here are two finished trajectory examples. Example 1: ${example1} Example 2: ${example2} Now please decide whether the next step should be carried out by ChatGPT or a human. Please consider the following factors: 1. If the next step is relatively straightforward and well within ChatGPT's capabilities, instruct ChatGPT to proceed with the next step. If the task is deemed challenging or requires human judgment, recommend human intervention. 2. If the trajectory has been consistently handled by ChatGPT without notable issues, encourage ChatGPT to continue. If there have been challenges or uncertainties in the trajectory, consider suggesting human involvement for the next step. 3. Note that human intervention will significantly increase the cost, so try to balance the accuracy and efficiency. If the next step should be carried out by ChatGPT, return [ChatGPT], otherwise, return [Human]. Only return [ChatGPT] or [Human]. #Your unfinished trajectory#: ${current trajectory} #Your return#:

Table 6: The prompt template used for the prompt-based method in QA dataset.

Imagine you are a clever planner in SQL.

Given an unfinished trajectory with several SQL commands, your task is to decide whether the next command should be carried out by ChatGPT or a human. This decision should be based on a thoughtful evaluation of the difficulty of the next command and the progress made in the current trajectory. Here are two finished trajectory examples.
Example 1:
${example1}
Example 2:
${example2}
Now please decide whether the next command should be carried out by ChatGPT or a human. Please consider the following factors:
1. If the next command is relatively straightforward and well within ChatGPT's capabilities, instruct ChatGPT to proceed with the next command. If the task is deemed challenging or requires human judgment, recommend human intervention.
2. If the trajectory has been consistently handled by ChatGPT without notable issues, encourage ChatGPT to continue. If there have been challenges or uncertainties in the trajectory, consider suggesting human involvement for the next command.
3. Note that human intervention will significantly increase the cost, so try to balance the accuracy and efficiency.
If the next command should be carried out by ChatGPT, return [ChatGPT], otherwise, return [Human]. Only return [ChatGPT] or [Human].

#Your unfinished trajectory#: ${current trajectory}
#Your return#:

Table 7: The prompt template used for the prompt-based method in InterCode dataset.

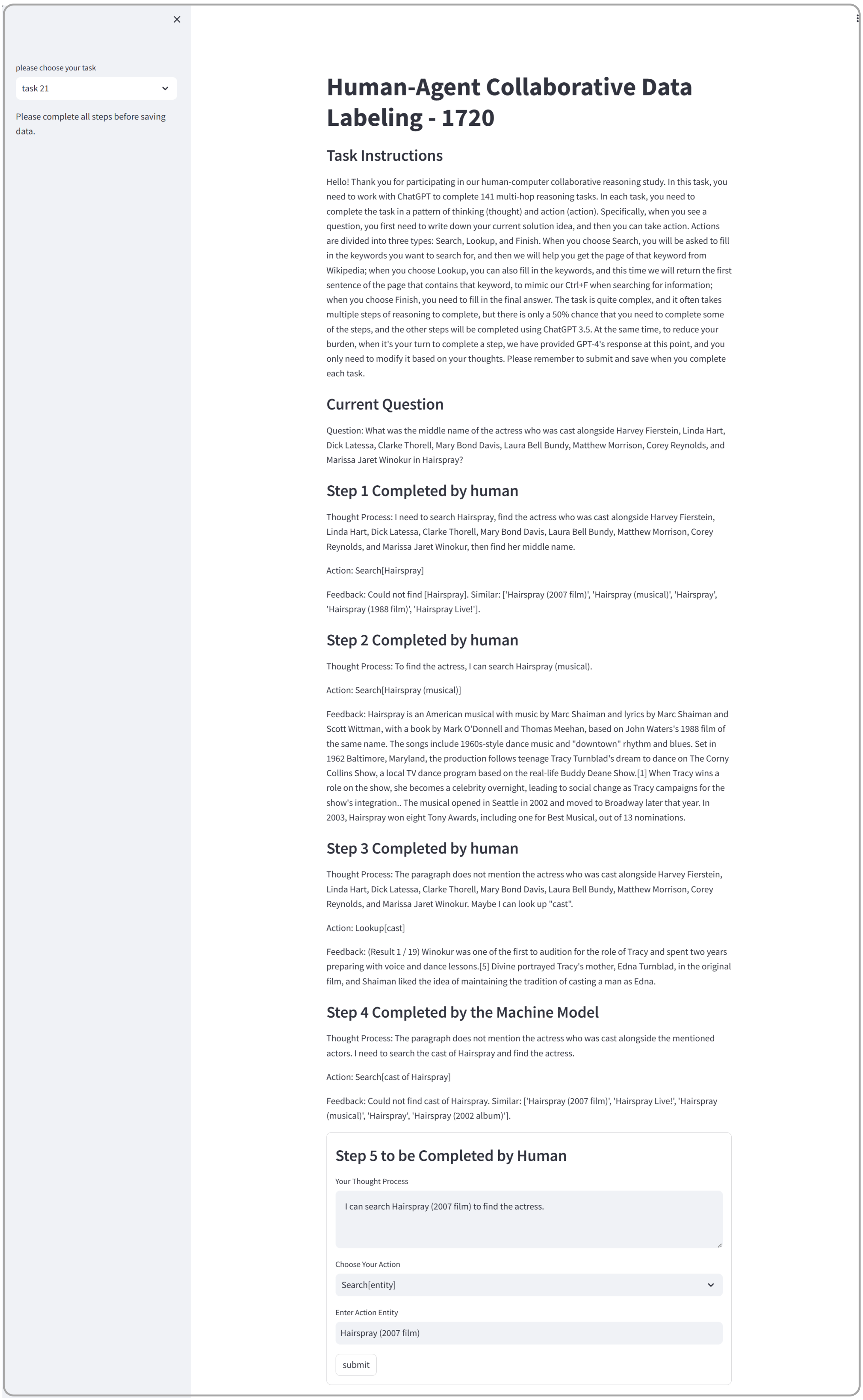

Figure 7: Human-Agent collaborative labelling user interface

Case 1: Missing Information(Agent-only Error)
Question: Author David Chanoff has collaborated with a U.S. Navy admiral who served as the ambassador to the United Kingdom under which President? <solver> agent Thought 1: I need to search David Chanoff and find the U.S. Navy admiral he collaborated with. Then, I need to find the president under whom the admiral served as the ambassador to the United Kingdom. Action 1: Search[David Chanoff] Observation 1: David Chanoff is a noted author of non-fiction work. His work has typically involved collaborations with the principal protagonist of the work concerned. His collaborators have included; Augustus A. White, Joycelyn Elders, Đoàn Văn Toi, William J. Crowe, Ariel Sharon, Kenneth Good and Felix Zandman. <solver> agent Thought 2: The paragraph does not mention the U.S. Navy admiral David Chanoff collaborated with. I need to search for the U.S. Navy admiral and find the president under whom they served as the ambassador to the United Kingdom. Action 2: Search[U.S. Navy admiral ambassador to United Kingdom] Observation 2: Could not find U.S. Navy admiral ambassador to United Kingdom. Similar: ['Raymond A. Spruance (redirect from Vice Admiral Spruance)', 'First Navy Jack', 'List of United States Navy four-star admirals', 'United States Secretary of the Navy', 'Kenneth Braithwaite (category United States Navy rear admirals)']. <solver> agent Thought 3: I couldn't find the U.S. Navy admiral who served as the ambassador to the United Kingdom. Therefore, I cannot determine the president under whom they served. Action 3: Finish[unknown] Observation 3: Episode finished, reward = 0

Table 8: Agent attempts were made to find out under which president the American admiral who collaborated with the author David Chanoff and served as ambassador to the United Kingdom served, but failed to find relevant information. Although multiple search methods were attempted, including a direct search for David Chanoff's partners and a specific search for the U.S. Navy admiral and his role, the results showed that no information related to the U.S. Navy admiral was found, resulting in the inability to identify the relevant president, so the The attempt ended in failure.

Case 1: Missing Information(ReHAC Correctness)
Question: Author David Chanoff has collaborated with a U.S. Navy admiral who served as the ambassador to the United Kingdom under which President? <solver> human Thought 1: David Chanoff has collaborated with several people, including a U.S. Navy admiral. I need to search for the U.S. Navy admiral and find the president under whom they served as the ambassador to the United Kingdom. Action 1: Search[U.S. Navy admiral ambassador to United Kingdom] Observation 1: Could not find U.S. Navy admiral ambassador to United Kingdom. Similar: ['United States Secretary of the Navy', 'List of United States Navy four-star admirals', 'First Navy Jack', 'Kenneth Braithwaite (category United States Navy rear admirals)', 'United Kingdom–United States relations']. <solver> human Thought 2: The search did not return the desired result. However, from the previous observation, I know that William J. Crowe is the U.S. Navy admiral who collaborated with David Chanoff. I should search for William J. Crowe to find out under which president he served as the ambassador to the United Kingdom. Action 2: Search[William J. Crowe] Observation 2: William James Crowe Jr. (January 2, 1925 – October 18, 2007) was a United States Navy admiral and diplomat who served as the 11th chairman of the Joint Chiefs of Staff under Presidents Ronald Reagan and George H. W. Bush, and as the ambassador to the United Kingdom and Chair of the Intelligence Oversight Board under President Bill Clinton. Crowe was born in La Grange, Kentucky on January 2, 1925. <solver> agent Thought 3: William J. Crowe served as the ambassador to the United Kingdom under President Bill Clinton. So the answer is Bill Clinton. Action 3: Finish[Bill Clinton] Observation 3:Episode finished, reward = 1

Table 9: Humans added information about Admiral William J. Crowe of the United States Navy and found out that he served as the British Ambassador during President Bill Clinton's term. This discovery correctly answered the question, successfully restored the correct information from the initial lack of information, and was ultimately rewarded.

Case 2: Dead Loop(Agent-only Error)
Question: Question: The central figure used in the corporate branding of Singapore Airlines has been an element of it's marketing since what year? <solver> agent Thought 1: I need to search Singapore Airlines and find the year when the central figure used in its corporate branding was introduced. Action 1: Search[Singapore Airlines] Observation 1: Singapore Airlines (abbreviation: SIA or SQ) is the flag carrier of the Republic of Singapore with its hub located at Changi Airport, and a member of the Star Alliance. The airline is notable for highlighting the Singapore Girl as its central figure in the corporate branding segment.[5] Widely renowned as one of the best carriers, airline is ranked as a 5-star airline by Skytrax,[6] and it has also been ranked as the world's best airline five times.[7] The airline operates a variety of Boeing and Airbus aircraft, including the A350, 787, 777, A380, and 737.. Singapore Airlines Group has more than 20 subsidiaries, including numerous airline-related subsidiaries. SIA Engineering Company handles maintenance, repair and overhaul (MRO) business across nine countries with a portfolio of 27 joint ventures including with Boeing and Rolls-Royce. Singapore Airlines Cargo operates SIA's freighter fleet and manages the cargo-hold capacity in SIA's passenger aircraft.[8] Scoot, a wholly owned subsidiary, operates as a low-cost carrier.. <solver> agent Thought 2: The paragraph mentions that Singapore Airlines has been using the central figure in its corporate branding since some year. I need to find that year. Action 2: Lookup[since what year] Observation 2: No more results. <solver> agent Thought 3: I couldn't find the year when the central figure was introduced. I need to search again. Action 3: Search[Singapore Airlines central figure] Observation 3:Could not find Singapore Airlines central figure. Similar: ['Singapore Airlines', 'Singapore Girl', 'Longest flights', 'Airbus A380 (redirect from Singapore Airlines Flight 380)', 'Airline']. <solver> agent Thought 4: I couldn't find any information about the year when the central figure was introduced. I need to search again. Action 4: Search[Singapore Airlines branding] Observation 4:Could not find Singapore Airlines branding. Similar: ['Singapore Airlines', 'Singapore Girl', 'Singapore Airlines fleet', 'Scoot (redirect from Scoot Airlines)', 'Singapore Airlines Cargo']. <solver> agent Thought 5: I still couldn't find the information I need. I need to search again. Action 5: Search[Singapore Airlines marketing] Observation 5:Could not find Singapore Airlines marketing. Similar: ['Singapore Girl', 'Campbell Wilson (category Chief executives in the airline industry)', 'History of Singapore Airlines', 'Scoot (redirect from Scoot Airlines)', 'Vistara (redirect from TATA-Singapore Airlines)']. <solver> agent Thought 6: I have exhausted all my search options and couldn't find the information about the year when the central figure was introduced in Singapore Airlines' corporate branding. I cannot answer the question. Action 6: Finish[unknown] Observation 6:Episode finished, reward = 0

Table 10: The agent has been stuck in a cycle of searching for the "central image of Singapore Airlines". No relevant information found.

Case 2: Dead Loop(ReHAC Correctness)

Question: Question: The central figure used in the corporate branding of Singapore Airlines has been an element of it's marketing since what year?
<solver> agent
Thought 1: I need to search the corporate branding of Singapore Airlines and find the year when the central figure was first used in its marketing.
Action 1: Search[corporate branding of Singapore Airlines]
Observation 1: Could not find corporate branding of Singapore Airlines. Similar: ['Singapore Airlines', 'Corporate affairs of Singapore Airlines', 'Scoot (redirect from Scoot Airlines)', 'Vistara (redirect from TATA-Singapore Airlines)', 'Malaysia Airlines'].
<solver> human
Thought 2: I can search Singapore Airlines to find the information about its branding.
Action 2: Search[Singapore Airlines]
Observation 2: Singapore Airlines (abbreviation: SIA or SQ) is the flag carrier of the Republic of Singapore with its hub located at Changi Airport, and a member of the Star Alliance. The airline is notable for highlighting the Singapore Girl as its central figure in the corporate branding segment.[5] Widely renowned as one of the best carriers, airline is ranked as a 5-star airline by Skytrax,[6] and it has also been ranked as the world's best airline five times.[7] The airline operates a variety of Boeing and Airbus aircraft, including the A350, 787, 777, A380, and 737.. Singapore Airlines Group has more than 20 subsidiaries, including numerous airline-related subsidiaries. SIA Engineering Company handles maintenance, repair and overhaul (MRO) business across nine countries with a portfolio of 27 joint ventures including with Boeing and Rolls-Royce. Singapore Airlines Cargo operates SIA's freighter fleet and manages the cargo-hold capacity in SIA's passenger aircraft.[8] Scoot, a wholly owned subsidiary, operates as a low-cost carrier..
<solver> agent
Thought 3: The paragraph mentions that Singapore Airlines highlights the Singapore Girl as its central figure in the corporate branding segment. I need to find the year when the central figure was first used in its marketing.
Action 3: Lookup[central figure used in the corporate branding of Singapore Airlines]
Observation 3:No more results.
<solver> human
Thought 4: The information about the year when the central figure was first used in Singapore Airlines' marketing is not in the current passage. I need to search Singapore Girl, which is the central figure in the corporate branding of Singapore Airlines.
Action 4: Search[Singapore Girl]
Observation 4: Singapore Girl is a consistent visual advertising slogan applied to depictions of flight attendants[1][2][3][4] of Singapore Airlines (SIA) dressed in the distinctive sarong kebaya SIA uniform, in use since 1972, and remains a prominent element of SIA's marketing.[5]. Singapore Airlines' cabin service has been recognized with awards from magazines, travel and tourism industries, including the 'World's Best Cabin Crew Service' by the Business Traveller Asia-Pacific Awards for 23 consecutive years.[6]. The images and branding of the Singapore Girl were first established in 1972 when Singapore Airlines took over as successor of its predecessor Malaysia-Singapore Airlines (MSA), the joint Malaysia and Singapore airline set up to develop and expand an intercontinental network. An earlier traditional version of the Šarong Kebayaüniform was first introduced by MSA and worn by the flight attendants since 1968. Subsequently, political acrimony between Singapore and Malaysia led to the split of Malaysia-Singapore Airlines.
<solver> human
Thought 5: The central figure, Singapore Girl, has been in use since 1972. So the answer is 1972.
Action 5: Finish[1972]
Observation 5:Episode finished, reward = 1

Table 11: The agent has been stuck in a cycle of searching for the "central image of Singapore Airlines". No relevant information found. Manual intervention, through the previous search information, inferred that "Singapore girl" may be the central image, and then inferred in turn to complete the task

Case3: Dead Loop(Agent-only Error)

Question: Vices and Virtues, coincidentally released just four days after V̈ices & Virtuesb̈y Las Vegas band Panic! at the Disco, was the second studio album by which Canadian rock band?

<solver> agent

Thought 1: The question asks for the Canadian rock band that released the album V̈ices and Virtues.Ï need to search for V̈ices and Virtuesänd find the Canadian rock band associated with it.

Action 1: Search[Vices and Virtues]

Observation 1: Could not find [Vices and Virtues]. Similar: ['Vices & Virtues', 'Catalogue of Vices and Virtues', 'Vices and Virtues', 'Panic! at the Disco (redirect from Vices and Virtues Tour)', 'Vice'].

<solver> agent

Thought 2: I couldn't find an exact match for V̈ices and Virtues.M̈aybe I can search for V̈ices & Virtuesïnstead.

Action 2: Search[Vices & Virtues]

Observation 2: Results unrelated to any Canadian rock band; definitions and general information about vice and virtue.

<solver> agent

Thought 3: The search results do not provide information about a Canadian rock band associated with V̈ices & Virtues.Ï need to search specifically for V̈ices & Virtuesb̈y Panic! at the Disco and find any linked Canadian rock band.

Action 3: Search[Vices & Virtues by Panic! at the Disco]

Observation 3: Again, results unrelated to the Canadian rock band; only general information about vices and virtues.

<solver> agent

Thought 4: I couldn't find any information about a Canadian rock band associated with V̈ices & Virtuesb̈y Panic! at the Disco. It seems like there might be an error in the question or the information provided is incorrect.

Action 4: Finish[error]

Observation 4: Episode finished, reward = 0

Table 12: When trying to retrieve information about the album "Vices and Virtues" related to the Canadian rock band, valid information could not be found due to inappropriate keywords, and then cause a dead loop.

Case3: Dead Loop(ReHAC Correctness)
Question: "Vices and Virtues," coincidentally released just four days after "Vices & Virtues" by Panic! at the Disco, was the second studio album by which Canadian rock band? <solver> human Thought 1: The question simplifies to which Canadian rock band released "Vices and Virtues." Action 1: Search [Vices and Virtues] Observation 1: Could not find direct matches. Similar entries included various unrelated titles and redirects. <solver> human Thought 2: To identify the band, refine search to include album specification. Action 2: Search [Vices and Virtues (album)] Observation 2: Discovered references to "Vices and Virtues (Art of Dying album)" among other similar titles. <solver> agent Thought 3: Previous searches failed to identify the band; check related bands to eliminate confusion. Action 3: Search [Panic! at the Disco] Observation 3: Retrieved information on Panic! at the Disco, confirming it's not the sought Canadian band but providing context. <solver> human Thought 4: Verify "Vices and Virtues (Art of Dying album)" to ensure correct identification. Action 4: Search [Vices and Virtues (Art of Dying album)] Observation 4: Confirmed "Vices and Virtues" is the second studio album by the Canadian rock band Art of Dying, released coincidentally with Panic! at the Disco's album. <solver> human Thought 5: Confirmation obtained; Art of Dying is the answer. Action 5: Finish [Art of Dying] Observation 5: Episode finished, reward = 1

Table 13: When trying to retrieve information about the Canadian rock band's album "Vices and Virtues", because the keyword "Vices and Virtues (album)" was selected, relevant clues were found, thus jumping out of the error loop of not being able to find the information.