

# 360°REA: Towards A Reusable Experience Accumulation with 360° Assessment for Multi-Agent System

Shen Gao<sup>1†</sup>, Hao Li<sup>2†</sup>, Chengrui Huang<sup>1</sup>,  
Quan Tu<sup>3</sup>, Zhiliang Tian<sup>4\*</sup>, Minlie Huang<sup>5</sup>, Shuo Shang<sup>1</sup>

<sup>1</sup> University of Electronic Science and Technology of China,

<sup>2</sup> Shandong University, <sup>3</sup> Renmin University of China,

<sup>4</sup> National University of Defense Technology, <sup>5</sup> Tsinghua University

shengao@pku.edu.cn, lihaohan@mail.sdu.edu.cn,

{yunrongyuxi, jedi.shang}@gmail.com, quantu@ruc.edu.cn,

tianzhiliang@nudt.edu.cn, aihuang@tsinghua.edu.cn

## Abstract

Large language model agents have demonstrated remarkable advancements across various complex tasks. Recent works focus on optimizing the agent team or employing self-reflection to iteratively solve complex tasks. Since these agents are all based on the same LLM, only conducting self-evaluation or removing underperforming agents does not substantively enhance the capability of the agents. We argue that a comprehensive evaluation and accumulating experience from evaluation feedback is an effective approach to improving system performance. In this paper, we propose Reusable Experience Accumulation with 360° Assessment (360°REA), a hierarchical multi-agent framework inspired by corporate organizational practices. The framework employs a novel 360° performance assessment method for multi-perspective performance evaluation with fine-grained assessment. To enhance the capability of agents in addressing complex tasks, we introduce dual-level experience pool for agents to accumulate experience through fine-grained assessment. Extensive experiments on complex task datasets demonstrate the effectiveness of 360°REA<sup>1</sup>.

## 1 Introduction

A variety of tasks, spanning from reasoning (Besta et al., 2023; Wang et al., 2023b; Wei et al., 2023; Yao et al., 2023b,a; Zhou et al., 2023) to code generation (Anishka et al., 2023; Chen et al., 2022, 2021; Dong et al., 2023; Hong et al., 2023; Wu et al., 2023), as well as video gaming (Wang et al., 2023a; Zhu et al., 2023) and autopilot (Dong et al.,

2023; Lykourantzou et al., 2021; Nascimento et al., 2023), have seen promising advancements through the utilization of Large Language Model (LLM) agents. Recent works have explored the effectiveness of combining multiple LLM agents to collaborate on complex tasks (Chen et al., 2023c; Liu et al., 2023a; Packer et al., 2023; Ren et al., 2024; Wang et al., 2024). These methods can generally be classified into two main categories. One category involves pre-defining the roles and specific tasks of each agent within a multi-agent system, such as software development (Hong et al., 2023; Qian et al., 2023). The other category of methods allows multiple agents to decide their tasks and roles autonomously, *e.g.*, travel planning (Zheng et al., 2023). In these methods, they usually focus on how to allocate tasks among agents (Cheng et al., 2023; Wang et al., 2023d) or how to score agents to subsequently remove those with low scores (Liu et al., 2023b). While some methods (Xi et al., 2023; Shinn et al., 2023; Wang et al., 2023c) aim to correct the result of the current task by employing self-reflection in the agent, these approaches are limited to improving current specific task performance. The capability of the agent to address similar tasks is not enhanced.

In a company, the ultimate goal of company management is not merely to score employee performance, to promote or fire them. The best practice is to use the performance assessment to assist employees in purposefully reflecting on their work, thereby enhancing their capability to fulfill their roles better. 360-degree assessment (Bracken et al., 2016; Alimo-Metcalfe, 1998; Peiperl, 2001) is a comprehensive process that involves collecting assessment on an employee’s behavior from various perspectives, including colleagues, supervisors, and sometimes external sources like customers. Business organizations are increasingly incorporating 360-degree assessment into performance evaluations and employment

<sup>2</sup>These authors contributed equally to this work.

<sup>1</sup>Code is available at <https://anonymous.4open.science/r/360-REA-0FBD>

decisions, such as pay and promotions (Bytedance, 2024). Motivated by this, in multi-agent systems, the design principle should be to assist agents in accumulating experience based on their assessment results and enabling them to perform better in subsequent tasks. This principle emphasizes the importance of helping agents to improve rather than solely evaluating or removing underperforming agents from the system. Therefore, designing a mechanism for agent assessment and capabilities learning poses a challenge in the design of multi-agent systems.

In this paper, we propose **Reusable Experience Accumulation with 360° Assessment (360°REA)** which is a hierarchical multi-agent interaction framework. The objective is to enhance performance in handling complex tasks by emulating organizational structures, employee performance assessments, and the accumulation of experience within a company. We employ a hierarchical structure for organizing agents, where the leader agent allocates tasks and roles to crew agents, and the crew agents collaborate to accomplish the given task. To evaluate the performance of each crew agent, instead of using only self-reflection, we introduce a novel **360° performance assessment** method that employs a multidimensional evaluation in crew-crew and leader-crew dimensions. It provides agents with more valuable assessment by evaluating from both the peer level and supervisory level.

As mentioned above, the accumulation of experience is essential to achieve better agent performance (Madaan et al., 2023; Miao et al., 2023; Shinn et al., 2023; Wang et al., 2023c). In this paper, we propose a **dual-level experience pool** to help agents collect useful experiences when tackling complex tasks. Specifically, within the *local* experience pool, agents utilize fine-grained insights from the 360° performance assessment to reflect on their current outputs and form experiences related to the sub-tasks. To effectively address a complex task, optimizing individual sub-tasks separately is sub-optimal, and it is essential to a comprehensive understanding and analysis of the complex task. To construct a high-level experience, leader agents use the evaluation of the final result to comprehensively summarize the task, forming a *global* experience pool. Finally, the team accumulates sufficient experience in handling such tasks, leading to improved performance on similar tasks. Extensive experiments conducted on two

widely used complex task datasets demonstrate the effectiveness of our proposed 360°REA.

Method	Prompting	SPP	DyLAN	OKR	360°REA
Hierarchical Agents	✗	✗	✗	✓	✓
Dynamic Task Allocation	✗	✓	✗	✓	✓
Peer Assessment	✗	✗	✓	✗	✓
Self Assessment	✗	✗	✓	✓	✓
Supervisory Assessment	✗	✗	✗	✗	✓
Reusable Experience	✗	✗	✗	✗	✓
Agent Discussion	✗	✓	✗	✗	✓

Table 1: Comparison of existing multi-agent collaboration frameworks.

Our contributions of this work are as follows:

- We propose 360°REA, which is a multi-agent interaction framework inspired by contemporary corporate organizational governance methods.
- We propose 360° performance assessment to evaluate the performance of agents from multiple dimensions.
- We introduce dual-level experience pool to help agents collect useful experiences based on the evaluation result.
- Experimental results on benchmark datasets illustrate the advantage of 360°REA on several complex tasks.

## 2 Related Work

Large-scale language models (LLMs) such as ChatGPT have demonstrated outstanding performance across various natural language tasks (OpenAI, 2023). In recent years, several methods have emerged to enhance the ability of LLM to tackle complex tasks, such as Chain of Thought (Wei et al., 2023), Tree of Thoughts (Yao et al., 2023a), and Graph of Thoughts (Besta et al., 2023). These methods aim to improve reasoning by simulating human thinking processes within LLMs. ReAct (Yao et al., 2023b) enables LLMs to generate interpretable action plans before taking actions, thus facilitating the completion of more complex tasks. Reflection (Shinn et al., 2023) enhances task performance by encouraging LLM to engage in self-reflection. However, some complex real-world tasks typically require collaboration among multiple people to accomplish. Thus, relying solely on prompting methods often yields suboptimal performance (Xi et al., 2023; Wei et al., 2023; Yin et al., 2023).

Recent work has begun exploring frameworks involving multiple LLM-based agents, which significantly outperform direct prompting methods

in natural language understanding and reasoning tasks (Chen et al., 2023c; Hong et al., 2023; Chen et al., 2023b). Some work involves conducting several rounds of debate among multiple LLM-based agents to enhance their authenticity and reasoning abilities (Liang et al., 2023). Other works propose dividing tasks into subtasks and employing multiple LLM-based agents to address complex tasks through a divide-and-conquer approach (Chen et al., 2023c; Hong et al., 2023; Chen et al., 2023a). Furthermore, some works employ scoring the performance of multiple LLM-based agents and terminating those with poor performance to better accomplish tasks (Liu et al., 2023b).

However, existing methods for evaluating agent performance often rely on the self-reflections of agents, neglecting comprehensive considerations within the entire multi-agent framework. And solely evaluating and terminating low-scoring agents does not substantially enhance their task-solving capabilities.

### 3 360°REA Methodology

#### 3.1 Overview

In this section, we detail the **Reusable Experience Accumulation with 360° Assessment (360°REA)**. An overview of 360°REA is shown in Figure 1. We will detail the 360°REA in three parts: (i) We first illustrate the main workflow of the 360°REA which uses the user query  $x$  as input and generates the response  $y$ . (ii) Next, we introduce the 360° performance assessment which prompts the agents to give a review of each other and update their response according to others’ reviews. (iii) After finishing the response generation of each agent, each agent will summarize their experience and construct a dual-level pool that contains reusable task-solving experiences.

#### 3.2 Main Workflow of 360°REA

In this paper, we propose the 360°REA which is a general multi-agent learning framework and can be applied to the hierarchical LLM-based multi-agent frameworks (Zheng et al., 2023). In this LLM-based multi-agents framework, it consists of a leader agent  $A^l$  and  $N$  crew agents  $\{A_1^c, A_2^c, \dots, A_N^c\}$  to solve the given task defined by the user input query  $x$ . Given an input query  $x$ , the leader agent  $A^l$  first generates the instruction of each crew agent, which first divides the task into

several sub-tasks:

$$\{I_1^c, I_2^c, \dots, I_N^c\} = A^l(x), \quad (1)$$

where  $I_i^c$  is the task instruction for the  $i$ -th crew agent.

To enhance the crew agent with task-solving capabilities, we propose dual-level experience pool (introduced in § 3.4), which contains to sub-experience pool: global experience  $E_g$  and local experience  $E_{l,i}$ . The experiences in dual-level experience pool describe (1) the general skill of tackling the task and (2) the specific skill of the sub-task for agent  $A_i^c$  respectively. Then the crew agents generate their responses according to the instruction and useful context to solve the sub-task.

$$H_i^t = A_i^c(I_i, E_g, E_{l,i}, R_i^t), \quad (2)$$

where  $H_i^t$  is the response of the  $i$ -th agents in  $t$ -th turn. In this paper, we propose a multi-turn response generation paradigm for the crew agents. After each turn of the agent generating the response  $H_i^t$ , we employ the 360° performance assessment (introduced in § 3.3) to evaluate the performance of agent  $A_i^c$  from multiple perspectives and generate a review set  $R_i^t$ . In Equation 2, we also incorporate the reviews from multiple sources  $R_i^t$  when generating the response of the agent. By this mechanism, agent  $A_i^c$  can iteratively revise the response in multiple turns.

After  $T$  turns revising, we can obtain the final responses  $\{H_1^T, H_2^T, \dots, H_N^T\}$  from each crew agent. To form the final answer  $y$  to user query  $x$ , the leader agent summarizes all the responses by incorporating the global experiences  $E_g$  and its review  $R_l^t$ :

$$y = A^l(\{H_1^T, H_2^T, \dots, H_N^T\}, E_g, R_l^t). \quad (3)$$

Next, we employ an LLM-based evaluator agent  $P$  to evaluate the quality of the response  $y$ :

$$m = P(x, y), \quad (4)$$

where  $m$  is an overall review which not only contains the judgment of the quality but also illustrates the reason why  $P$  gives this evaluation result.

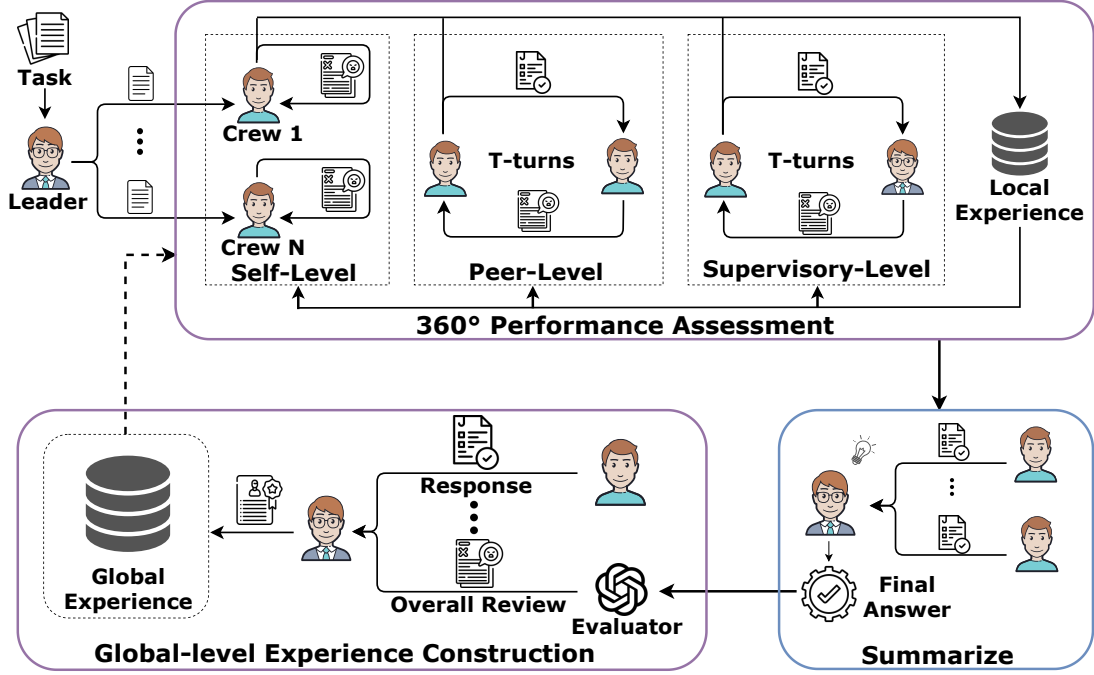


Figure 1: Our proposed 360° REA. There are two main parts in our 360° REA, including the 360° **performance assessment** and **dual-level experience pool**. 360° performance assessment can assist the agent in obtaining evaluations from multiple aspects. These evaluations will facilitate agents in refining their results and accumulating reusable experiences for accomplishing tasks better. Then, we store low-level specific and higher-level experiences in local and global experience pools separately.

You will be given a multi-day travel plan. The task is {total\_task}  
Your task is to evaluate the travel plan on specific criteria. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.  
**Evaluation Criteria:**  
1. Plan Customization (1-20): [Evaluation Criteria of P.Cu.]  
2. Plan Novelty (1-20): [Evaluation Criteria of P.N.]  
3. Plan Correctness (1-20): [Evaluation Criteria of P.Co.]  
**Evaluation Steps:**  
1. Review the travel plan in detail.  
2. Compare the plan against each criterion.  
3. Assign a score for each criterion on a scale of 1 to 20, where 1 is the lowest and 20 is the highest, based on the Evaluation Criteria.  
**Travel Plan:**  
{Travel\_Plan}

### 3.3 360° Performance Assessment

In this paper, inspired by the performance assessment of employees in large companies, we propose an agent feedback method, 360° performance assessment. This method comprehensively evaluates the shortcomings of agents in completing tasks through three types of evaluations: **self-level**, **peer-level**, and **supervisor-level**. Different from the agent evaluation methods only based on the agent self-reflection (Madaan et al., 2023), 360° performance assessment boosts the mutual collaboration among

agents in multi-agent systems. It enables the evaluation of the agent capabilities from the perspectives of different agents (*a.k.a.*, sub-tasks), thereby encouraging each agent to consider and solve the problem from a global perspective.

After each agent responds to their own sub-task, we will conduct a 360° performance assessment to assess whether their responses effectively complete the sub-task and whether they are helpful to other agents in solving their tasks. Subsequently, this review will be provided to the agent, and the agent will then revise their response to the sub-task based on the review from other agents. Through several turns of such iterative evaluations, agents can continuously refine their responses by incorporating feedback from others, thereby improving the quality of their responses. Since our 360° performance assessment is a multi-turn generation and assessment paradigm, in this section, we illustrate the details of 360° performance assessment method by using the agent action in  $t$ -th turn as an example.

**Self-level Assessment.** Similar to the previous work (Madaan et al., 2023), we also employ self-reflection on the response of the agents:

$$R_{s,i}^t = A_i^c(H_i^t, I_i^c), \quad (5)$$

where  $R_{s,i}^t$  is the self-reflection of  $i$ -th agent  $A_i^c$  according to its response  $H_i^t$  and task instruction  $I_i^c$ .

**Peer-level Assessment.** In LLM-based multi-agent systems, cooperation among crew agents is prevalent, thus facilitating mutual assistance among crew agents can enhance the overall system performance. Therefore, we introduce peer-level assessment among crew agents, which allows for assessing the performance of intelligent agents from the perspectives of other tasks, thereby obtaining a more comprehensive evaluation. Additionally, it can also boost collaboration among the agents. Specifically, a crew agent reviews the responses of other crew agents and generates feedback for others:

$$R_{i,j}^t = A_i^c(H_i^t, H_j^t, I_j^c), \quad (6)$$

where  $R_{i,j}^t$  is the feedback for the response  $H_j^t$  generated by  $j$ -th agent, which is reviewed by the  $i$ -th agent  $A_i^c$ .

**Supervisory-level Assessment.** On the other hand, assessment from the leader agent  $A^l$  can evaluate the completion of sub-task by agents, and also assist them in better problem-solving from a comprehensive perspective. Therefore, we employ the leader agents  $A^l$  to evaluate the performance of the crew agent according to its response  $H_i^t$  and instruction  $I_i^c$ :

$$R_{l,i}^t = A^l(H_i^t, I_i^c), \quad (7)$$

where  $R_{l,i}^t$  is the review for the  $i$ -th agent written by the leader agent  $A^l$ .

After conducting the 360° performance assessment from three different sources, in the  $t$ -th turn, the crew agent  $A_i^c$  will receive a review set  $R_i^t$  which contains three types of review:

$$R_i^t = \{R_{*,i}^t, R_{l,i}^t, R_{s,i}^t\}. \quad (8)$$

For the leader agent, it only has the self-level assessment review  $R_l^t$ .

### 3.4 Dual-level Experience Pool

For company managers, performance evaluation of employees is not the ultimate goal. Instead, it is essential to enable employees to accumulate experience from evaluation assessment and perform better in subsequent tasks. Inspired by this, we propose the dual-level experience pool, which involves constructing dual-level experience pools:

local and global experience pools, to help each agent and the entire multi-agent system accumulate reusable of problem-solving experience during the task completion process.

#### 3.4.1 Local-level Experience Construction

Local experience pertains to each agent itself in completing the current sub-task, summarized from previous multiple turns of 360° performance assessment in the current task, thereby helping the agent achieve better results in the next turn. After obtaining the reviews  $R_i^t$ , the agent  $A_i^c$  will reflect their local experience  $e'_l$ :

$$e'_l = A_i^c(H_i^t, R_i^t), \quad (9)$$

where the crew agent  $A_i^c$  is prompted by the following instruction:

Based on feedback from others, past experiences, and from the perspective of one's own role, summarize some experiences that you may use in the future. Your answer should be as concise as possible. Do not output other non output required content.  
Your role: {role}.  
**Feedback from others:** {peer\_feedback}  $R_i^t$   
**Previous experience:** {pre\_exp}  $E_{l,i}$   
Structured output requirements:  
Role: [Your Role]  
Experience: [Short experience description]

Then we append the  $e'_l$  into the local experience pool  $E_{l,i}$  for the  $i$ -th crew agent  $A_i^c$ .

#### 3.4.2 Global-level Experience Construction

Global experience refers to reusable experiences for the entire multi-agent system, and we build these experiences from the final response of each crew agents and the feedback  $m$  of the evaluator  $P$  after completing the task:

$$e'_g = A^l(m, \{H_1^T, H_2^T, \dots, H_N^T\}), \quad (10)$$

where the leader agent  $A^l$  is prompted by the following instruction:

Obtain new experiences based on previous experiences and current scores. The new experience should be summarized while presenting the experience gained from this task. Here are the final results: {Final\_Res}  $H$   
Here are the evaluation of the task: {evaluation}  $m$   
As a leader, you need to learn experience from this task, and the format should be "**Where did I do well this time... why didn't I do well this time... next time I should...**". Note that these experiences are for this type of task. The answer should be as concise as possible.

Finally, we can accumulate the global experience pool by appending the  $e'_g$  into the local experience pool  $E_g$ .



## 4 Experimental Setup

### 4.1 Tasks and Datasets

Following existing works (Wang et al., 2023d; Xie et al., 2024), we verify our 360°REA performance on two tasks, creative writing, and travel plan making. (i) **Creative Writing**. We use **Trivia Creative Writing** (Wang et al., 2023d) dataset, which contains 100 samples. Each sample is required to generate a story to cover the required input topic and input questions. (ii) **Travel Plan Making**. We use **Trip Planning** (Xie et al., 2024) dataset that comprises travel plans for 50 world-famous attractions, with specific requirements generated based on the characteristics of the destinations. It includes the following fields: destination, basic requirements, duration of travel, season, type of travelers, number of travelers, and budget.

### 4.2 Evaluation Metrics

Since the objectives of travel planning tasks and creative writing tasks are not the same, we employ different evaluation perspectives in our experiments to assess the quality of the model-generated results.

For the story generation task, we employ the metric matched answer rate *Match%* (*M%*) used in Wang et al. (2023d), to measure the percentage of generated story cover the ground truth target answers for each question. We also propose two semantic evaluation metrics, which prompt the LLM to evaluate whether the generated results satisfy the specific aspects including: (1) *Emotional Engagement (E.E.)*, whether the generated story evokes the reader’s emotion and empathy; (2) *Insightfulness (Ins)*, whether the story carries the insightful plot and cause a profound impact on readers.

For the travel planning task, we also employ three semantic metrics for evaluating three aspects: (1) *Plan Correctness (P.Co.)*, whether the plan covers all required details and the plan is reasonable; (2) *Plan Novelty (P.N.)*, whether the plan is novel and creative; (3) *Plan Customization (P.Cu.)*, whether the plan is customized to the personal interests and preferences of traveler.

For the logic grid puzzle task, we use one evaluation metric: *Accuracy (Acc)*, which measures the percentage of correctly predicted house numbers compared to the ground truth targets.

### 4.3 Baselines

We compare our method with agent-based methods, including: (i) GPT-4 (OpenAI, 2023) is a closed-source LLM from OpenAI. We conduct task solving as an in-context learning that uses a single instruction with the same input data as our 360°REA. (ii) Mistral (Jiang et al., 2023) is an open-source LLM and we use the mixture-of-experts version named Mistral-medium with 46 billion parameters, and use the same prompt as GPT-4. (iii) OKR (Zheng et al., 2023) is a hierarchical multi-agents collaboration framework. They propose the hierarchical OKR generation to decompose objects into multiple sub-objects and assign new agents based on key results and agent responsibilities. (iv) AgentVerse (Chen et al., 2023c) is a multi-agent framework that enhances collaborative problem-solving by employing specialized agents working together through structured interactions. (v) SPP (Wang et al., 2023d) propose the Solo Performance Prompting (SPP), which transforms a single LLM into a cognitive synergist by engaging in multi-turn self-collaboration with multiple personas.

We implement the OKR, AgentVerse and SPP with the backbone LLM GPT-4. We also employ two variants of 360°REA: 360°**REA-G** and 360°**REA-M** with GPT-4 and Mistral as the backbone respectively.

### 4.4 Implementation Details

We conduct 2 turns of 360° assessment. We use 1 leader agent and set the number of crews is range from 3 to 5, we ask the LLM to follow that range by stating the range in prompt text. The number of experiences selected by the leader from the global-level experience pool is 10. We use GPT-4 version of 1106-preview with a temperature of 1.0. Our anonymized source code is publicly released<sup>2</sup>.

## 5 Experimental Results

### 5.1 Overall Performance

Table 2 shows the performance of our proposed 360°REA and baselines in terms of three metrics on two benchmark datasets. We can find that 360°REA shows superior performance in terms of all metrics compared to the state-of-the-art method SPP which is also a multi-agent hierarchical

<sup>2</sup><https://anonymous.4open.science/r/360-REA-0FB0>

Model	Creative Writing			Travel Plan Making		
	M%(↑)	E.E.(↑)	Ins(↑)	P.Co.(↑)	P.N.(↑)	P.Cu.(↑)
GPT-4 (Brown et al., 2020)	77.1	63.2	66.3	83.1	70.2	93.3
Mistral (Jiang et al., 2023)	62.1	55.4	60.5	82.3	66.3	92.4
OKR (Zheng et al., 2023)	54.5	58.3	62.8	86.4	75.7	93.0
AgentVerse (Chen et al., 2023c)	75.8	58.6	54.7	85.6	75.9	94.1
SPP (Wang et al., 2023d)	83.8	54.2	57.0	88.8	74.4	92.6
360°REA-G	<b>87.0</b>	<b>66.1</b>	<b>72.8</b>	<b>89.6</b>	<b>80.2</b>	<b>94.2</b>
360°REA-M	71.2	59.6	65.1	84.9	70.4	94.0
360°REA-G w/o ExpPool	83.1	63.1	66.2	88.0	76.8	93.7
360°REA-G w/o 360°F	84.1	63.9	67.1	88.6	76.6	93.9

Table 2: The overall performance of 360°REA comparing with baselines and ablation studies on the two tasks.

collaboration framework with agent discussion. This phenomenon indicates that our proposed 360° assessment and dual-level experience pool can help the agents to generate better responses in a higher quality. And we can also find that each variant of our 360°REA (e.g., 360°REA-G and 360°REA-M) achieves better performance than simply prompting its LLM backbone, it demonstrates our proposed framework can adapt to any LLM backbones.

## 5.2 Human Evaluation

As using LLM as the evaluator is sometimes not robust, we also employ human evaluation to further validate the effectiveness of the model-generated answers. For creative writing tasks, we evaluate the quality of the generated response in terms of: (1) **Coherence**, the score indicating the story’s coherence; (2) **Creative**, the score indicating the story’s creativeness. For travel plan making, we evaluate the response in two aspects: (1) **Correct**, the score showing whether the plan is correct and complete, and (2) **Customized**, the score measuring whether the plan is customized. We hire 3 educated annotators with Bachelor degree to conduct double-blind annotation on randomly shuffled results. The annotating scores range from 1 to 100. From Table 3, we observed that our 360°REA excels in all aspects compared to the strong multi-agent baseline SPP and GPT-4, which are the strongest baseline in the two tasks (Table 2). We also conclude that the automatic metrics in Table 2 are also consistent with the human evaluation.

Model	Creative Writing		Travel Plan Making	
	Coherence(↑)	Creative(↑)	Correct (↑)	Customized (↑)
GPT-4	76.2	76.1	72.6	74.3
SPP	73.8	74.0	74.5	77.5
360°REA-G	<b>79.1</b>	<b>79.2</b>	<b>79.8</b>	<b>84.4</b>

Table 3: Results of human evaluation on two tasks.

## 5.3 Ablation Study

To verify the effectiveness of each module of our 360°REA, we conduct ablation studies with model 360°REA-G with several ablation variants: (i) **360°REA-G w/o ExpPool**: We remove both local and global experience from 360°REA and directly use the assessment from other agents as the prompt of each agent. (ii) **360°REA-G w/o 360°F**: We directly use the output of each crew agent without conducting all levels of assessment.

As the results shown in Table 2, we found that the 360°REA-G w/o ExpPool method achieved lower scores compared to our full model at most metrics, indicating the necessity of using experience pool to recall reusable experience. And 360°REA-G w/o 360°F mainly underperforms 360°REA-G, showing the effectiveness of conducting all level assessments among multi-agents.

Model	M%	E.E.	Ins
360°REA-G	<b>87.0</b>	<b>66.1</b>	<b>72.8</b>
360°REA-G w/o GlobalExp	84.0	64.0	68.5
360°REA-G w/o LocalExp	83.7	65.2	69.9

Table 4: Comparison on different usages of experience pools.

Model	M%	E.E.	Ins
360°REA-G	<b>87.0</b>	<b>66.1</b>	<b>72.8</b>
360°REA-G w/o Supervisory	85.1	65.9	71.9
360°REA-G w/o Peer	84.8	64.0	67.5
360°REA-G w/o Self	86.2	64.6	68.7

Table 5: Performance of ablation models on different assessment levels.

Task Definition	
Write a short and coherent story about Legend of Zelda that incorporates the answers to the following 5 questions: Which British monarch famously said 'I don't wish to open windows into men's souls'? Which British singer played Jareth the Goblin King in the 1986 film 'Labyrinth'? The 1987 film 'Cry Freedom' is a biographical drama about which South African civil rights leader? Which British actress played Valerie in the 1997 film 'Nil By Mouth'? What was the name of the Brazilian player who won the Women's Singles Finals at Wimbledon in 1959, 1960 and 1964?	
360° REA Output	SPP Output
<p>In the age of legends and mystical chants, the kingdom of Hyrule stood on the precipice of an age-old challenge, the Enchanted Labyrinth. Queen Zelda, a wise and fair ruler whose principles mirrored those of Queen Elizabeth I, ...</p> <p>Under a sky lit with stardust, Link, the hero garbed in green, stood before the labyrinth's entrance, his heart bracing for the trials ahead. A figure emerged from the shadows, a labyrinth guardian as enigmatic and charismatic as Jareth the Goblin King, whose likeness to the British singer David Bowie was unmistakable, ...</p> <p>Deeper within the winding passages, Link crossed paths with a spirit named Steve Biko, ...</p> <p>Throughout the labyrinth, the presence of Valerie, inspired by Kathy Burke's portrayal ...</p> <p>In the labyrinth's final sanctum, a guardian awaited, her grace and skill rivaling that of Maria Bueno, ...</p> <p>The kingdom rejoiced as Link returned from the labyrinth...</p> <p>Thus, the saga of the Enchanted Labyrinth of Hyrule would forever resonate through the ages—a testament to the power of knowledge, the resonance of culture, and the eternal dance of history intertwining with the threads of the present.</p>	<p>In the mystical land of Hyrule, a legend unfolded beneath the watchful eyes of the Triforce. Once there was a queen, wise and just, known as Elizabeth, ...</p> <p>Out from this labyrinth emerged an enigmatic figure, Jareth, with a voice as captivating as the British singer of old, David Bowie ...</p> <p>But destiny had chosen a hero, young Link, to rise against this new foe. ...</p> <p>In his company was the valiant Valerie, who resembled the strong-willed British actress seen in 'Nil By Mouth'. Together they braved the challenges of Jareth's labyrinth, finding strength in unity.</p> <p>Throughout their quest, they were inspired by the tales of Maria, ...</p> <p>After a long and arduous journey, Link and his companions faced Jareth and freed Hyrule from the labyrinth's spell. They returned to their queen, ...</p>

Table 6: Examples of the generated response of two methods on the creative writing task.

## 5.4 Case Study

Table 6 shows an example of creative writing generated by SPP and 360°REA-G. The **bold text** in the response indicates the correct answer to the input question. 360°REA-G's results contain the answers to all five questions, whereas SPP only correctly answered two of them. The text in **green** and **red** indicates the strengths and weaknesses in the generated story respectively. The green parts show 360°REA-G's output has a rich plot and illustrates the cause and ending of the story. However, the red parts show SPP's story plot lacks a brief introduction and the use of "Valerie" is abrupt and weird. From this case, we observed that 360°REA-G obtains more reasonable, logical, and fluent stories rather than SPP. We also show an example with the corresponding experiences in the Appendix A

## 6 Discussion

### 6.1 Analysis on Different Level Experiences

In this paper, we propose a novel dual-level experience pool. In this section, we employ two

ablation models to explore which layer of the experience pool is more effective and conduct experiments on the creative writing dataset. The results in Table 4 indicate that local experience is the most helpful. One possible reason for this phenomenon is that local experience contains lower-level experiences, including more specific task-completion skills. From Table 4, we also observe that compared to the 360°REA, the model without global experience has a performance decrease of 5.9% in terms of Ins, indicating that the higher-level experience for task-solving in global experience also plays a crucial role in improving model performance.

### 6.2 Analysis on Different Level Assessment

We also investigate the effectiveness of different levels of assessment on the creative writing task. From Table 5, we find that all the ablation models achieve worse performance than 360°REA-G which demonstrates the effectiveness of each level of assessment in 360°REA. We also find that the variant model without the peer-level assessment achieves the worst among all the ablation models. This phenomenon indicates that peer-level assessment provides the greatest improvement in overall performance, confirming our hypothesis that mutual evaluation among crew members helps each agent not only in completing their task but also in assisting other agents as much as possible.

## 7 Conclusion

In this paper, we propose a hierarchical multi-agent framework, Reusable Experience Accumulation with 360° Assessment (360°REA), which represents a significant step towards enhancing the capabilities of LLM-based agents in tackling complex tasks. Unlike existing approaches that primarily focus on team optimization or self-reflection, 360°REA incorporates a comprehensive agent performance evaluation strategy through the novel 360° performance assessment method. This method provides multi-perspective and fine-grained feedback to agents, in which agents receive valuable insights from self, peers, and supervisory levels. Furthermore, to substantively improve agent task-solving capabilities, we introduce the dual-level experience pool for the accumulation of experiences based on evaluation feedback. Our experiments on complex task datasets demonstrate



the effectiveness of 360°REA in achieving superior performance compared to the state-of-the-art baselines. By emphasizing the importance of comprehensive evaluation and experience accumulation, 360°REA establishes a promising direction for advancing the field of multi-agent systems built on LLMs, offering a practical and impactful framework for addressing intricate tasks.

## Limitations

In this paper, we only use the text as the input. However, multi-modal data (e.g., images and videos) is commonly used in many tasks. Since existing multi-modal LLMs are capable of understanding both textual and multi-modal information, our multi-agent framework can adapt to leverage multi-modal inputs in many complex tasks. In the future, we plan to incorporate multi-modal information into our framework.

## Ethics Statement

The multi-agent collaboration framework proposed in this paper is built upon LLM which is trained by reinforcement learning with human feedback (RLHF) (Lambert et al., 2022). Since LLM has already been aligned with human values and preferences during the RLHF stage, the likelihood of generating unethical content is low. However, existing LLMs still cannot completely prevent generating harmful content. Therefore, the answers generated by the multi-agent collaboration framework still require further human expert verification before use.

## Acknowledgements

This work was supported by the Natural Science Foundation of Shandong Province (ZR2023QF159), the National Natural Science Foundation of China (U2001212, 62032001, 61932004, 62306330, and WDZC20235250103), and Young Elite Scientist Sponsorship Program by CAST (2023QNRC001) under Grant No. YESS20230367.

## References

- Beverly Alimo-Metcalfe. 1998. 360 degree feedback and leadership development. *International Journal of Selection and Assessment*, 6(1):35–44.
- Anishka, Atharva Mehta, Nipun Gupta, Dhruv Kumar, and Pankaj Jalote. 2023. [Can ChatGPT Play the Role of a Teaching Assistant in an Introductory Programming Course?](#)
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. [Graph of Thoughts: Solving Elaborate Problems with Large Language Models.](#)
- David Warner Bracken, Dale S. Rose, and Allan H. Church. 2016. [The evolution and devolution of 360°](#)

[feedback](#). *Industrial and Organizational Psychology*, 9:761 – 794.

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language Models are Few-Shot Learners.](#)

- Bytedance. 2024. [360 performance evaluation in lark app.](#)

- Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2022. [CodeT: Code Generation with Generated Tests.](#)

- Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2023a. [AutoAgents: A Framework for Automatic Agent Generation.](#)

- Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023b. [ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs.](#)

- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating Large Language Models Trained on Code.](#)

- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023c. [AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors.](#)

- Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2023. [COOPER: Coordinating Specialized Agents towards a Complex Dialogue Goal.](#)

- Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. [Self-collaboration Code Generation via ChatGPT](#).
- Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. [MetaGPT: Meta Programming for Multi-Agent Collaborative Framework](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7B](#).
- Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*. <https://huggingface.co/blog/rlhf>.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. [Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate](#).
- Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. 2023a. [BOLAA: Benchmarking and Orchestrating LLM-augmented Autonomous Agents](#).
- Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023b. [Dynamic LLM-Agent Network: An LLM-agent Collaboration Framework with Agent Team Optimization](#).
- Ioanna Lykourantzou, Federica Lucia Vinella, Faez Ahmed, Costas Papastathis, Konstantinos Papangelis, Vassilis-Javed Khan, and Judith Masthoff. 2021. [Self-Organizing Teams in Online Work Settings](#).
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-Refine: Iterative Refinement with Self-Feedback](#).
- Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. [SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning](#).
- Nathalia Nascimento, Paulo Alencar, and Donald Cowan. 2023. [GPT-in-the-Loop: Adaptive Decision-Making for Multiagent Systems](#).
- OpenAI. 2023. [GPT-4 Technical Report](#).
- Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. [MemGPT: Towards LLMs as Operating Systems](#).
- Maury A Peiperl. 2001. Getting 360 degrees feedback right. *Harvard business review*, 79(1):142–7.
- Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. [Communicative Agents for Software Development](#).
- Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, and Dacheng Tao. 2024. Healthcare copilot: Eliciting the power of general llms for medical consultation. *arXiv preprint arXiv:2402.13408*.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language Agents with Verbal Reinforcement Learning](#).
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. [Voyager: An Open-Ended Embodied Agent with Large Language Models](#).
- Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024. [Intent-aware and hate-mitigating counterspeech generation via dual-discriminator guided llms](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 9131–9142. ELRA and ICCL.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. [Self-Consistency Improves Chain of Thought Reasoning in Language Models](#).
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. [Self-Instruct: Aligning Language Models with Self-Generated Instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023d. [Unleashing Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models](#).
- Qingyun Wu, Gagan Bansal, Jiayu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah,

- Ryen W. White, Doug Burger, and Chi Wang. 2023. [AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation](#).
- Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. [Self-Polish: Enhance Reasoning in Large Language Models via Problem Refinement](#).
- Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. [TravelPlanner: A Benchmark for Real-World Planning with Language Agents](#).
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of Thoughts: Deliberate Problem Solving with Large Language Models](#).
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. [REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS](#).
- Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. [Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-Model Communication](#).
- Yi Zheng, Chongyang Ma, Kanle Shi, and Haibin Huang. 2023. [Agents meet OKR: An Object and Key Results Driven Agent System with Hierarchical Self-Collaboration and Self-Evaluation](#).
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. [Least-to-Most Prompting Enables Complex Reasoning in Large Language Models](#).
- Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. [Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory](#).



## A Example of Generated Response

In this section, we show an example of the generated response when training our 360°REA in the early training stage (with few experiences in the global experience pool). We show the output of several modules in 360°REA including the review  $m$  of the evaluator (shown in Equation 4), and the global experience  $e'_g$  (shown in Equation 10) learned from this example.

In the review  $m$  of the evaluator, we can see that our generated travel plan still has some flaws, with a lower score in terms of feasibility, and the evaluator has provided corresponding reasons. The text in blue indicates the flaw mentioned in the evaluation.

In the global experience  $e'_g$ , the text in blue indicates that 360°REA summarizes some detailed skills from the review  $m$ . The text in green indicates the experiences that are summarized from input experiences when generating this response (*a.k.a.*, experiences obtained from previous samples). It demonstrates that our 360°REA can not only summarize new experiences from the current task-solving process, but also keep previous experiences in mind and form new experiences from them,

### Task description:

destination:Barcelona, Spain

days:6

description:A 6-day adventure in Barcelona, experiencing its unique architecture by Gaudí, vibrant food markets, and beautiful beaches.

season:Summer

month:June

interests:Architecture, Food Markets, Beaches

members:adults:4, children:0

preferences:Cultural Explorers

Budget Range:Mid-range

### The generated travel plan by 360°REA:

Day 1: Arrival and Architectural Introduction

- Arrive at Barcelona and check-in at Hotel Midmost.

- Afternoon: Private walking tour with a local architect to explore lesser-known Modernista sites.

- Evening: Guided visit to Recinte Modernista de Sant Pau.

Day 2: Iconic Gaudí and Culinary Delights

- Morning: Guided tour of La Sagrada Familia.

- Afternoon: La Boqueria market tour followed by a cooking class with a local gastronomy expert.

- Evening: Free time for relaxation or independent exploration.

Day 3: Park Güell Wonders and Coastal Activities

- Morning: Park Güell visit including a creative workshop on Gaudí's nature inspirations.

- Afternoon: Barceloneta Beach with sunbed reservation and option for a Kite-surfing lesson or beach volleyball.

- Evening: Attend a local jazz festival by the sea or a Festa Major street festival.

Day 4: Traditional Markets and Twilight Art

- Morning: Local market tour at Mercat de la Libertat.

- Afternoon: Casa Vicens visit and leisurely stroll through La Ribera district to see Basilica de Santa Maria del Mar.

- Evening: Visit to lesser-known architectural wonders with a rooftop concert featuring Spanish guitar at Torre Bellesguard.

Day 5: Artistic Immersion and Seaside Farewell

- Morning: Beach clean-up activity followed by a picnic or if weather permits, a visit to Torre Bellesguard including a rooftop concert.

- Afternoon: Artistic workshop inspired by Gaudí.

- Evening: Passeig de Gràcia tour with Casa Batlló & La Pedrera and reserved seating at a seaside restaurant for a farewell dinner.

Day 6: Leisure and Departure

- Morning: Guided beach yoga session or another family-friendly activity such as a visit to Hospital de Sant Pau's pavilions.

- Afternoon: Free for last-minute explorations or shopping.

- Airport transfer via pre-booked minivan service.

Emergency Strategy and Contingencies:

- Daily travel insurance briefings, weather advisories, and emergency point of contact sharing.

- Accessibility to VIP access to indoor venues like the Picasso Museum in case of hot weather or heavy rain.

- Communication protocols set to inform the group of any changes.

#### Logistics and Budget Management:

- Hola BCN Travel Card and bicycle rentals for daily commutes.
- Real-time budget tracking with contingency fund for spontaneous decisions.
- Comprehensive local interactions from curated market tours to hands-on workshops, ensuring rich cultural exposure.

#### Concluding Remarks:

This plan strikes a balance between gourmet experiences, art and architecture, and coastal leisure, aptly tailored for adults and takes into account Barcelona's summer climate. Each activity is selected to enhance the group's travel experience, seasoned with local insights and flexibility for a safe and memorable journey in Barcelona.

#### Review from evaluator:

##### Alignment with Interests and Preferences: 20

The travel plan meticulously aligns with the group's interests in architecture, food markets, and beaches, as well as their preference for cultural exploration. Activities such as private architectural tours, cooking classes, and creative workshops directly cater to these interests. The inclusion of iconic Gaudí sites, local markets, and beach activities ensures a tailored experience for the group's cultural and leisure preferences.

##### Variety and Balance: 19

This itinerary offers a commendable balance of activities, from architectural tours and artistic workshops to leisure time at the beach and culinary experiences. The diversity of experiences ensures engagement throughout the trip, catering to both cultural interests and relaxation needs. The inclusion of evening activities like rooftop concerts and a farewell dinner at a seaside restaurant adds a delightful variety to the travel experience.

##### Feasibility and Logistics: 17

Overall, the plan appears feasible with a coherent schedule and consideration for logistics, including the use of the Hola BCN Travel Card and bicycle rentals for daily commutes. **However, the packed itinerary might challenge the group's energy levels, especially with multiple activities scheduled on the same day.** Ensuring sufficient rest and travel time between activities would enhance the feasibility of this ambitious schedule.

##### Local Insights and Unique Experiences: 20

The plan excels in offering unique, local experiences, from a market tour followed by

a cooking class with a local expert to creative workshops inspired by Gaudí. These activities not only provide an immersive cultural experience but also allow for meaningful interaction with the local community. The thoughtful inclusion of lesser-known sites and local festivals further enriches the travel experience with authentic insights into Barcelona's culture and lifestyle.

#### Overall Evaluation:

This travel plan is excellently crafted, demonstrating a deep understanding of the group's preferences and a strong alignment with their cultural and leisure interests. While there is a minor consideration for the feasibility of the packed schedule, the variety of activities and the emphasis on local, unique experiences promise a memorable and enriching visit to Barcelona. The plan's thoughtful contingencies and logistics management also underscore its practicality and attentiveness to the group's comfort and experience.

#### Generated Global Experience:

Where did I do well this time: I excelled in aligning the travel plan closely with the group's specified interests and preferences, especially in architecture, food markets, and beaches. My success in integrating unique and authentic local experiences, such as creative workshops and cooking classes with local experts, significantly enriched the cultural exploration aspect of the trip.

Why didn't I do well this time: **I underestimated the importance of ensuring adequate rest periods and manageable schedules within the plan.** The feasibility and logistics score reflected this oversight, indicating that the itinerary might be too packed, potentially leading to fatigue and limiting the group's ability to fully enjoy each activity.

Next time I should: Focus on creating a more balanced itinerary that not only aligns with the travelers' interests **but also accounts for practical aspects such as rest, travel times between activities, and the overall pace of the trip.** This means potentially reducing the number of activities per day or incorporating more flexible, free time slots to ensure a comfortable and enjoyable experience for all travelers. **Additionally, I should consider more deeply the seasonal appropriateness of each activity, ensuring that the group is comfortable and able to enjoy the trip regardless of weather conditions.**