ACT: Knowledgeable Agents to Design and Perform Complex Tasks

Makoto Nakatsuji[1], **Shuhei Tateishi**[2], **Yasuhiro Fujiwara**[3],
Ayaka Matsumoto[1], **Narichika Nomoto**[1], **and Yoshihide Sato**[1]
[1]NTT Human Informatics Laboratories
[2]NTT DOCOMO, INC.
[3]NTT Communication Science Laboratories
makoto.nakatsuji@ntt.com

Abstract

Large language models enhance collaborative task execution in multi-agent systems. Current studies divide a complex task into manageable components for agents to solve. However, agents often lack a clear understanding of the overall task and each other's roles, hindering synergy and solution integration. We propose a method called knowledgeable ***A****gents* to design and perform ***C****omplex* ***T****asks* (ACT), where: (1) Agents independently manage their knowledge and tasks while collaboratively designing the complex task into a more comprehensible form. In parallel, each agent also acquires knowledge of others, defined as a structured description of how other agents approach their tasks based on the agent's own task resolution. (2) Each agent updates its knowledge and refines its task through interactions with others. By referencing structured knowledge, the agents effectively integrate their tasks to collaboratively solve the complex task. Three evaluations, including creative writing and tool utilization, show that ACT outperforms existing methods in terms of accuracy when solving complex tasks. Detailed prompt examples are included in the appendix to facilitate future research reuse.

1 Introduction

Large language models (LLMs), such as OpenAI's ChatGPT (OpenAI, 2023), Google's Gemini (Gemini-team et al., 2023), and Meta's LLaMA3 (Meta, 2024), have rapidly evolved and become pervasive in everyday human life. The advancements in these LLMs have drawn significant attention from researchers in developing collaborative LLM agents (Li et al., 2023a; Park et al., 2023). For example, some studies define the expertise of each agent and leverage it to extract domain knowledge from LLMs, which is then assigned to the agents. In a team composed of these agents, a complex task is divided into individual tasks, allowing each agent to generate and share solutions, thereby collaboratively solving the overall task (Chen et al., 2024c; Wang et al., 2023; Li et al., 2023a; Chen et al., 2024a; Qiao et al., 2024).

However, this approach has two limitations. First, the complex task assigned to the team is vague and lacks mechanisms for agents to clarify it. For example, the complex task in "Trivia Creative Writing" (Wang et al., 2023) merely links multiple heterogeneous tasks, assigning agents to solve the overall task[1]. Since not all agents grasp its intricate nature, accuracy tends to decline. Second, current methods lack a mechanism for each agent to acquire and utilize knowledge about how other agents address their respective tasks during collaboration. This prevents them from understanding the relationships between their individual tasks and the complex team task. Thus, agents struggle to integrate their individual solutions into a comprehensive one for the complex task.

Unlike AI agents, human teams naturally develop collaborative mechanisms when tackling complex tasks. While AI agents work independently on predefined tasks, human team members actively refine their understanding through discussions within and outside the team. This iterative exchange of knowledge enables them to dynamically adjust their approaches, ensuring that individual tasks align with the overall team objective. Such collaborative task design plays a crucial role in creative production, such as service product proposals and research policy decisions. By continuously reflecting each member's updated insights, these discussions enhance both the coherence and accuracy of the final outcomes (Hackman, 2002; Salas et al., 2008; Edmondson, 1999).

This study proposes a novel method, called

[1]They did not use "creative" to mean creativity or free-form writing as in WEAVER (Wang et al., 2024), but rather to assess how accurately the created text reflects the answer. We adopt the same usage of "creative writing."

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16831–16861

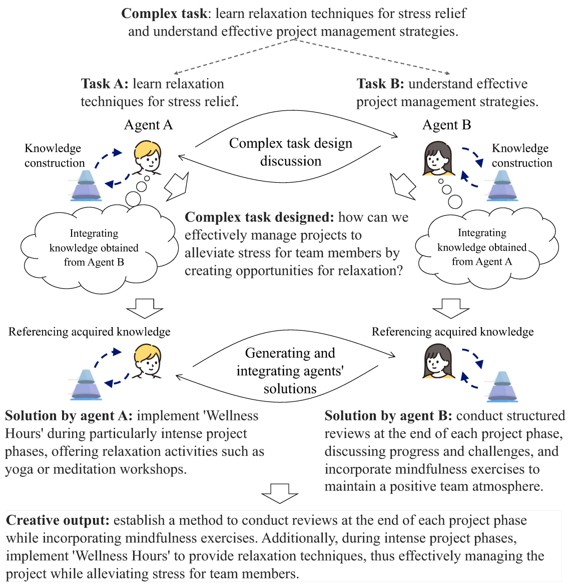


Figure 1: Image of creative output production in ACT.

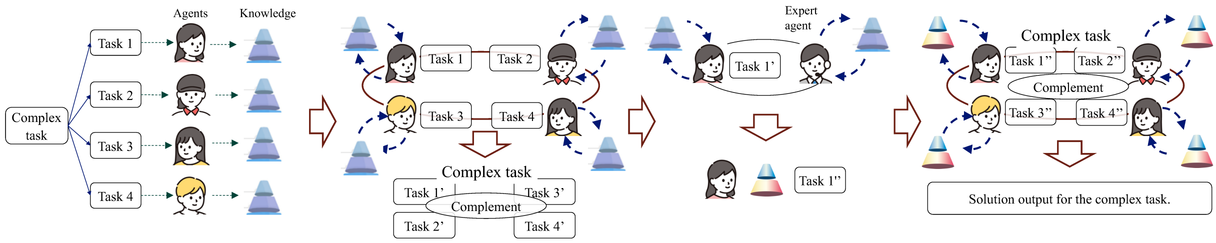
knowledgeable ***A***gents to design and perform ***C***omplex ***T***asks (ACT). ACT draws inspiration from approaches used in team collaboration within human society. Here, an agent's knowledge is defined as a structured description of how other agents approach their tasks from the perspective of the agent's own task resolution, linked with episode memories (Tulving, 2002), which are summarized representations of conversations with other agents in the past meetings. First, the initial complex task is distributed to each agent, who then acquires the necessary knowledge to solve their task and manages it independently. Next, agents collaboratively design the complex task, which is the process of aligning it with individual tasks while ensuring the entire team understands the complex task. Then, through repeated discussions with agents inside and outside the team, each agent gains insight into how their devised solution aligns with the overall team solution, facilitating its seamless integration into a cohesive outcome.

Let us illustrate this with the example in Figure 1. Suppose a complex task: "learn relaxation techniques for stress relief and understand effective project management strategies." Previous studies (Chen et al., 2024c; Wang et al., 2023) divide such a task into two independent parts, assigning one to Agent A and the other to Agent B. However, the lack of contextual clarity may lead to their solutions being merely concatenated without coherence. In contrast, if the task is framed as, "How can we manage projects to reduce team stress by creating relaxation opportunities?" it clarifies the connection between individual tasks and the overall objective. This alignment fosters a coherent relationship, ensuring that agents integrate their solutions meaningfully rather than treating tasks in isolation. Moreover, if Agents A and B exchange knowledge through discussions, they can leverage this shared knowledge to develop concrete solutions for the complex task. For example, Agent B can leverage the shared knowledge by utilizing insights about "stress reduction" acquired from Agent A to specify a solution: "To manage projects effectively, conduct structured reviews at the end of each project phase while incorporating mindfulness exercises to maintain a positive team atmosphere." By integrating agents' solution ideas, the overall approach becomes: "To effectively manage projects and alleviate stress on team members, we will conduct structured reviews and incorporate mindfulness exercises while also establishing 'Wellness Hours' during intense phases." Thus, designing, executing, and leveraging individual knowledge improve solution accuracy for complex tasks.

Our study explores solving complex tasks with heterogeneous subtasks by leveraging agents' cognitive synergies to identify diverse, accurate task-solution combinations. We thus conducted experiments using a non-factoid QA dataset from five Reddit communities, transformed into a task-solution dataset. Supporting multiple answers per question (Nakatsuji and Okui, 2020), it enables diverse task-solution evaluation. We also use the TriviaQA dataset (Joshi et al., 2017), offering limited answer diversity but used in prior work (Wang et al., 2023), and an open-ended QA dataset (Chen et al., 2024c), where agents generate solutions using tools like web search APIs or code execution environments. Following (Wang et al., 2023), we use an automatic metric to assess whether agents' task solutions are correctly included in creative outputs, instead of subjectively evaluating coherence or creativity. To ensure the reliability of this metric, we validate it through human consensus-based evaluations. Consequently, ACT significantly improved task resolution accuracy, outperforming existing methods.

2 Related work

As research on collaborative execution of complex tasks among agents (Li et al., 2023a; Du et al., 2024; Wu et al., 2023; Xu et al., 2023), Camel (Li et al., 2023a) breaks down a single prompt pro-



(a) Agent generation. (b) Team meeting. (c) Break time. (d) Production meeting.

Figure 2: Architecture of ACT.

vided by the user into a specific task and then implements a solution through role-playing by two agents to achieve that task. (Du et al., 2024) involves multiple LLM agents exchanging information and opinions on a specific topic to generate dialogue reports. However, they rely on manually designed agents and cannot design or perform complex tasks with acquired knowledge.

Recent studies have focused on dynamically generating agents (Wang et al., 2023; Chen et al., 2024c,a; Li et al., 2023c). SPP (Wang et al., 2023) aims to leverage cognitive synergy from LLMs by dynamically identifying and simulating different personas based on task inputs. (Chen et al., 2024a) highlights the reliability of generated agents and the execution of tasks via collaborative and self-improvement actions. AgentVerse (Chen et al., 2024c) generates execution plans through discussions among the generated agents and provides evaluation strategies for iterative execution. However, they lack mechanisms for agents to acquire task knowledge collaboratively and manage their own independently, hindering task alignment.

(Chen et al., 2024b; Li et al., 2023b) aim to improve complex-task solutions and higher-order reasoning abilities through multi-agent collaboration based on prompt optimization. However, they require few-shot learning, whereas ACT functions effectively with a zero-shot approach.

3 Method

Here, we detail our method, ACT.

3.1 Architecture

Figure 2 illustrates the structure of ACT. Its task solution procedure is as follows:

First, in the "Agent generation" phase, the complex task is divided into distinct tasks, designed to facilitate collaborative refinement. Then, agents are dynamically generated and subsequently assigned tasks, producing initial knowledge relevant to their tasks. Next, during the "Team meeting," agents engage in discussions based on their individual tasks and knowledge to design a complex task that the team needs to tackle. Specifically, through the meeting, each agent acquires knowledge about their individual tasks, as well as the knowledge of team members and the designed complex task. In the 'Break time", agents first analyze and refine their contributions to complex task design in the team meeting. They identify the knowledge they need to deepen to better contribute to the team's solution. To achieve this, they dynamically generate expert agents with the identified knowledge, discuss its implications, and further expand their understanding to explore tasks more effectively. Finally, in the "Production meeting," agents generate creative outputs aligned with the team's complex task based on the discussions and knowledge acquired so far, ensuring that the results reflect solutions of individual agent's tasks.

Below, we first define the tasks and knowledge before explaining each step in Figure 2.

3.2 Tasks and knowledge

Tasks ACT first clarifies the overall goal and scope of the complex task. It then decomposes the complex task into multiple distinct tasks, each defined around a key element to ensure orthogonality and minimal overlap. These tasks are assigned to individual agents such that each contributes complementarily to the overall solution. Mathematically, the complex task $\mathcal{C}$ is thus divided into component tasks $\mathcal{T}_{i,0}$, each assigned to a unique agent a_i, where the subscript 0 denotes the task's initial state. If j denotes the number of meetings held, agent a_i works on task $\mathcal{T}_{i,j}$ in meeting $\mathcal{D}_j$.

Each task includes a subject representing its core focus and a description outlining its objectives or required actions. These are presented in bullet-point format for clarity. Formally, the task assigned to agent a_i at meeting j is defined as:

$$\mathcal{T}_{i,j} = \{s_{i,j}, d_i\} \quad (1)$$

Here, $s_{i,j}$ indicates the agent's concrete design goal in the j-th meeting, while d_i contains static background information derived from the original complex task. Examples used in our evaluation are shown in Table 2.

Knowledge structure Previous study (Wang et al., 2023) indicates that an agent's expertise is deemed sufficient with a single sentence describing its role, which enables knowledge to be more flexibly extracted from LLMs and applied to various contexts. However, when solving complex tasks consisting of heterogeneous subtasks assigned to agents within a team, coordination becomes challenging. The domain knowledge derived from a one-sentence role description is often insufficient, making it difficult to integrate different tasks and achieve collaborative solutions.

Thus, ACT adopts two strategies: (1) First, the agent accumulates knowledge about the agents it interacts with from the perspective of its own task and applies this knowledge to complex task design and production activities. Specifically, we define which agents engage in the meeting and what knowledge they acquire from each other in a way that links it to episodic memory (Tulving, 2002), making it accessible for agents to utilize when collaborating on task design and production activities. (2) Additionally, while summarizing episodes into concrete knowledge, we prepare more abstracted knowledge to avoid overly stringent constraints on the domain knowledge extraction from the LLMs, thus enhancing the reusability of knowledge. These strategies enable agents to utilize knowledge acquired from team members for designing and solving complex tasks, leading to clearer task concretization and improved solution accuracy.

Let us formalize the knowledge following the strategies. The knowledge held by agent a_i is composed of a set of knowledge collections concerning L agents, including itself, denoted as $\{\mathbb{K}_{i,l}\}_{l=1}^{L}$. Below, we explain this structure using the knowledge collection of agent a_i regarding a_l as an example. $\mathbb{K}_{i,l}$ is represented as a collection of episode $\mathcal{E}_{i,l,j}$ and knowledge $\mathcal{K}_{i,l,j}$ as follows:

$$\mathbb{K}_{i,l} = \{\mathcal{E}_{i,l,j}, \mathcal{K}_{i,l,j}\}_{l=1}^{L} \tag{2}$$

The counter j indicates the episode or knowledge acquired during the j-th meeting $\mathcal{D}_j$.

The j-th episode $\mathcal{E}_{i,l,j}$ obtained when agent a_i has a meeting with another agent a_l. $\mathcal{E}_{i,l,j}$ contains information about which agent (for example, a_l) the episode is with, as well as an episode description, $E^m_{i,l,j}$, and is defined as follows:

$$\mathcal{E}_{i,l,j} = \{E^m_{i,l,j}, a_l\}_{m=1}^{M} \tag{3}$$

ACT extracts up to M episodes from a meeting for another agent a_l. When extracting $E^m_{i,l,j}$, ACT focuses on the speech of agent a_l and extracts segments within X words.

Then, ACT extracts knowledge from each episode. If we represent $\{B^n_{i,l,j}\}_{n=1}^{N}$ as a set of brief summaries created from episode $E^m_{i,l,j}$ and $\{K^k_{i,l,j}\}_{k=1}^{N_k}$ as a set of important keywords derived from the summaries, the structure of $\mathcal{K}_{i,l,j}$ is formulated as follows:

$$\mathcal{K}_{i,l,j} = \{(B^n_{i,l,j}, K^k_{i,l,j}), sc^n_{i,l,j}\}_{n=1,k=1}^{N,N^k} \tag{4}$$

Here, N^k is the upper limit of extracted keywords for a brief summary, and N is the upper limit of knowledge pairs in $\mathcal{K}_{i,l,j}$. The summary $B^n_{i,l,j}$ is structured as: "subject is definition with some supplementary explanations," representing knowledge summarized by the agent from the episode. The score $sc^n_{i,l,j}$ indicates the significance of that knowledge pair for agent a_i.

The knowledge collection $\mathbb{K}_{i,i}$ held by agent a_i has the same knowledge structure but includes unique specifications, as detailed in Section 3.3.

In practice, during the reflection phase, agents extract relevant information from the conversation history by identifying salient actions and interactions. Each episode includes a concise description of the target user's key behaviors. From these episodes, agents generate brief summaries, assign 2 to 4 relevant keywords, record the source, and compute an importance score (0–10) for future use. For a step-by-step example, see Table 24 in Appendix C.4.

3.3 Agent generation

As part of the initial task assignment $\mathcal{T}_{i,0}$, agent a_i is generated with initial knowledge derived from that task. Specifically, a set of keywords $\{K^{k=1}_{i,i,0}\}_{k}^{N_k}$ representing the expertise of agent a_i is generated, which is deemed useful for the agent's task resolution by an LLM prompt (see Appendix C.1). This expertise remains fixed (i.e., $\{K^{k=1}_{i,i,j}\}_{k}^{N_k} = \{K^{k=1}_{i,i,0}\}_{k}^{N_k}$) and is not associated with any specific episode (i.e., $\{B^n_{i,i,j}\}_{n=1}^{N} = \emptyset$). The initial knowledge $\mathcal{K}_{i,l,0}$ for other agents is empty. Examples of initial expertise are in Tables 10 and 11 (Appendix A.2).

3.4 Team meeting

Each agent shares its task, gathers opinions, and collaboratively designs a complex task.

3.4.1 Procedure

(1) The agent a_i explains how its task $\mathcal{T}_{i,j}$ contributes to the complex task $\mathcal{G}$ being designed.
(2) After hearing each agent's opinion, agents utilize their knowledge $\{\mathbb{K}_{i,l}\}_{l=1}^{L}$ to advise on refining the complex task $\mathcal{G}$, considering its feasibility and coherence.
(3) A leader is randomly selected from the team members. The team leader, taking into account the opinions and advice received, designs the team's complex task $\mathcal{G}$ from the perspective of "feasibility." While designing $\mathcal{G}$, the leader considers each agent's task $\mathcal{T}_{i,j}$ and is permitted to add supplementary information for fusing tasks rather than merely concatenating them, thereby organizing the overall task. It is formulated as:

$$\mathcal{G} = \mathrm{fuse}\,(\mathcal{T}_{1,j}; \ldots; \mathcal{T}_{i,j}; \ldots; \mathcal{T}_{L,j}) \qquad (5)$$

The function fuse fuses the tasks of each agent and supplementary information generated to integrate the tasks, forming the overall complex task $\mathcal{G}$ with feasibility in mind. It is implemented as an LLM prompt (see Appendix C.2).

Through re-designing the complex task, each agent gains insight into how others approach their respective tasks while also developing a more concrete and practical understanding of the overall task. This process ensures consistency across heterogeneous tasks, ultimately improving the accuracy of complex task execution.

3.4.2 Task and knowledge updates

After the team meeting or the subsequent break time (Section 3.5), each agent delves into its task and updates its knowledge.

Task exploration Task exploration is conducted based on agent a_i's task $\mathcal{T}_{i,j}$, initial task $\mathcal{T}_{i,0}$, meeting record $\mathcal{D}_j$, knowledge $\mathbb{K}_{i,i}$, and complex task $\mathcal{G}$, as represented as follows:

$$\mathcal{T}_{i,j+1} = f(\mathcal{T}_{i,j}, \mathcal{T}_{i,0}, \mathcal{D}_j, \mathbb{K}_{i,i}, \mathcal{G}) \qquad (6)$$

The function $f()$ is implemented as a prompt for the LLM (see Appendix C.3), allowing the agent to determine how to update its task in accordance with its initial task and the complex task based on the meeting record. The number of provided knowledge items is set to c, based on the importance score $sc^{n}_{i,l,j}$, while a_i's initial knowledge is always included. Examples of the task exploration can be found in Tables 10 and 11 in Appendix A.2.

Knowledge acquisition The agent a_i also refers to the meeting record and acquires knowledge $\mathcal{K}_{i,l,j}$ and the corresponding episode $\mathcal{E}_{i,l,j}$ based on the statements of speaker a_l. Specifically, it acquires knowledge based on the previous knowledge of both agents $\mathbb{K}_{i,i}$ and $\mathbb{K}_{i,l}$, its prior task, $\mathcal{T}_{i,j-1}$, and the initial task $\mathcal{T}_{i,0}$, as follows:

$$\mathcal{K}_{i,l,j}, \mathcal{E}_{i,l,j} = g(\mathcal{D}_j, \mathbb{K}_{i,i}, \mathbb{K}_{i,l}, \mathcal{T}_{i,j-1}, \mathcal{T}_{i,0}) \qquad (7)$$

The function $g()$ is implemented as an LLM prompt (see Appendix C.4). It evaluates how the other agent's knowledge $\mathcal{K}_{i,l,j}$ can support a_i's task and the team's complex task, assigning an importance score $sc^{n}_{i,l,j}$ to each knowledge pair. The number of provided knowledge items is set to c based on $sc^{n}_{i,l,j}$, while a_i's initial knowledge is always included. Examples of acquired knowledge are shown in Tables 10 and 11 (Appendix A.2).

3.5 Break time

To effectively manage both individual and complex team tasks, shared knowledge needs to be coherently organized. After the team meeting, agents engage in self-reflection to identify gaps in knowledge consistency between their own tasks and the team's complex task. To address these gaps, ACT dynamically generates domain expert agents and conducts one-on-one meetings to acquire the necessary knowledge. The procedure is:
(1) Agent a_i reflects on whether its task has contributed to the design of the complex task in the meeting. It considers which aspects are lacking in its knowledge to concretize its task and the complex task to contribute to the team. It then decides on a set of knowledge it wishes to deepen and formulates questions regarding that knowledge.
(2) a_i dynamically generates an expert agent a_d capable of answering such questions. Specifically, the "knowledge to be deepened" is set as the initial task $\mathcal{T}_{d,0}$ of agent a_d, which is created using the initial knowledge, as described in Section 3.3.
(3) Agent a_i poses questions about the knowledge it wishes to enhance to the domain expert agent a_d.
(4) Agent a_d utilizes its knowledge to respond to those questions.
(5) In response, the agent a_i considers whether there are any further questions, and if so, proceeds

to ask deeper questions.
(6) Steps (4) and (5) are repeated until agent a_i has no further questions or a certain number of turn-taking exchanges has been exceeded.
(7) Finally, agent a_i explores tasks and updates knowledge by using the procedures outlined in Section 3.4.2.

Examples of acquired knowledge in break time are shown in Tables 10 and 11 in Appendix A.2.

3.6 Production meeting

In the production meeting, the agents collaboratively produce creative outputs that include their solutions to the problems of the complex task designed in the team meeting. The steps are:
(1) Each agent a_i refers to its accumulated knowledge from past meetings as well as the knowledge of team members $\{\mathbb{K}_{i,l}\}_{l=1}^{L}$, and its own tasks $\mathcal{T}_{i,0}$ and $\mathcal{T}_{i,j}$ to detail its solution to be included in the creative output. The knowledge $\{\mathbb{K}_{i,l}\}_{l=1}^{L}$ from the perspective of agent a_i is referenced to ensure consistency with the tasks of other agents while formulating solutions for its own task to be incorporated into the output. The agent also ensures that its tasks $T_{i,0}$ and $T_{i,j}$ align with and contribute to the team's complex task G and outputs its assessment R_i as:

$$\mathcal{R}_i = h(\{\mathbb{K}_{i,l}\}_{l=1}^{L}, \mathcal{T}_{i,j}, \mathcal{T}_{i,0}, \mathcal{G}) \tag{8}$$

The function $h()$ enables the agent to express its opinion in a manner consistent with its task while also aligning with the complex task. It is implemented as a prompt for the LLM (refer to Appendix C.5). The number of knowledge items provided is set similarly to equation (7).
(2) If a task requires tool utilization, each agent a_i generates an instruction set for selecting and making LLM function calls from its opinion $\mathcal{R}_i$, tasks $\mathcal{T}_{i,j}$, and $\mathcal{T}_{i,0}$ using the function $\mathrm{instruct}()$. a_i then obtains tool execution results via function calling $\mathrm{call}()$ and updates its opinion as:

$$\mathcal{R}_i = \{\mathcal{R}_i, \mathrm{call}(\mathrm{instruct}(\mathcal{R}_i, \mathcal{T}_{i,j}, \mathcal{T}_{i,0}))\} \tag{9}$$

Function $\mathrm{instruct}()$ is implemented as LLM prompts (see Appendix C.6).
(3) The leader produces a draft of the creative output based on the opinions of the agents. During this process, the leader ensures that the output contains responses relevant to each agent's initial task $\mathcal{T}_{i,0}$ and sufficiently contributes to the team's complex task, while also verifying clarity and adherence to word count limitations.

Table 1: Statistics of our dataset.

	Tea	Coffee	Design	Archi	Fashion
Task counts	551	486	280	89	240
Detail counts	1,720	1,553	882	325	849
Solution counts	2,278	1,849	1,083	339	1,124
Task length	11.1	11.3	11.7	11.9	11.1
Detail length	15.4	14.7	13.8	15.2	15.8
Solution length	16.4	16.6	15.1	17.0	16.5

(4) After confirming that these conditions are met, the final output is determined. If the conditions are not satisfied, the leader identifies the issues and shares them within the team. Each agent will then reassert their opinions based on the issues. Steps (2) to (4) are repeated until all conditions are met.

4 Evaluation

This section evaluates ACT comprehensively.

4.1 Reddit creative writing

Dataset Human creative activities are diverse, leading to varied tasks and solutions. To reflect such real-world conditions, we used a non-factoid QA dataset with diverse tasks and solutions to evaluate ACT. This dataset includes 6,673 subject-answer pairs from five categories (Tea, Cafe, Design, Architecture, Fashion) extracted from Reddit (Henderson et al., 2019). Each triple (consisting of a subject, detailed supplements, and responses) was transformed into triples of tasks, details, and solutions using ChatGPT-4o. This allows us to quantitatively evaluate how agents with heterogeneous tasks collaborate to solve complex tasks within a diverse solution space. This task is more complex and challenging compared to Trivia Creative Writing (Section 4.2), which lacks answer diversity, and OpenQA (Section 4.3), where agents work on a single complex task. Table 1 presents the statistics of this dataset. Here, "Archi" stands for Architecture. Each length represents the word count for each entry, and each task may include one or more detail and solution entries. Table 2 provides examples from the dataset.

Methodology This evaluation follows the experimental design on Trivia creative writing outlined by (Wang et al., 2023). We set the team size to four agents and $c{=}20$ for knowledge items. For acquiring knowledge, $M{=}4$, $N{=}4$, $N_k{=}4$, and $X{=}40$ were used, as these parameters enhance accuracy. To ensure reproducibility, ACT was evaluated across five communities. For each of 100 iterations, tasks for four agents were randomly selected from the dataset to design and solve com-

Table 2: Example of subject, detailed descriptions (details), and solutions of a task in the tea dataset: each detail has multiple descriptions, and each task has several solutions, as distinguished by double quotation marks.

Subject	Details	Solutions
Finding teas that pair well with cream or almond milk for a creamy, sweet treat.	"The user has been enjoying iced Thai tea mixed with almond milk, heavy cream, and a pinch of Splenda," "The user typically prefers straight water teas but is looking for creamy combinations," "The user has plenty of floral and fruit teas but needs recommendations for teas that work well with milk and sweetener."	"Earl Grey is a classic tea to pair with milk," "Genmaicha and Masala Chai are good options," "A mix of Darjeeling and Assam can be used to create a Japanese-style Royal milk tea," "Lapsang Souchong pairs well with almond milk and can be sweetened if desired," "Masala Chai or breakfast blends are also good with milk," "Lavender Earl Grey with milk and sugar makes a great London Fog latte."

Table 3: Comparison of methods.

		Tea	Coffee	Design	Archi	Fashion
Camel	Rouge-L	45.47	42.25	44.05	39.69	41.44
	Rouge-1	48.04	44.46	46.81	42.20	43.72
	# of Words	435.52	430.62	435.54	424.79	435.58
SPP	Rouge-L	41.71	39.90	36.65	37.27	37.51
	Rouge-1	43.99	42.61	39.10	40.13	39.84
	# of Words	316.07	368.67	252.24	431.52	300.57
Agent Verse	Rouge-L	42.01	38.17	39.51	33.43	38.67
	Rouge-1	41.91	41.35	42.60	36.42	39.01
	# of Words	273.75	276.71	238.14	227.28	264.17
ACT	Rouge-L	**46.86**	**44.32**	**44.37**	**41.99**	**43.11**
	Rouge-1	**49.38**	**47.15**	**45.96**	***45.04***	**46.15**
	# of Words	416.96	417.78	412.51	432.89	409.99
ACT$^+$	Rouge-L	***48.02***	***46.04***	***46.02***	***42.10***	***43.89***
	Rouge-1	***50.94***	***48.93***	***48.54***	**44.81**	***46.95***
	# of Words	458.39	468.78	459.75	465.87	461.88

Table 4: LLM-judged method comparison.

	Tea	Coffee	Design	Archi	Fashion
Camel	2.87	2.97	3.49	2.61	3.08
SPP	3.28	3.58	3.55	3.18	3.50
AgentVerse	2.27	2.37	2.90	2.41	2.42
ACT	***3.57***	***3.69***	**3.57**	***3.36***	***3.69***

Table 5: Comparison of human evaluation.

	Camel	SPP	AgentVerse	ACT
Accuracy	2.26	1.91	1.75	**2.85**

plex tasks in teams. The results below are averages of these 100 iterations.

(Wang et al., 2023) measured accuracy by using the percentage of exact matches in output answers. However, exact match metrics are unsuitable for non-factoid solutions, which are typically longer than factoid ones (see Table 1). Instead, we used ROUGE (Lin, 2004), a common metric for text generation (Touvron et al., 2023; Hu et al., 2022), to assess accuracy. ROUGE measures the overlap between reference responses and generated texts. For agent team, accuracy is the average ROUGE score across individual agents' tasks. To account for multiple correct solutions per task, we use the highest ROUGE score. The generated outputs follow the marketing proposal framework, including the service name, target persona, executive summary, specific offerings, and revenue model (Osterwalder and Pigneur, 2010; Kotler et al., 2016).

Compared methods We compared ACT with the methods from Section 2: (1) Camel: two agents role-play to complete tasks in a step-by-step manner after the task specification. (2) SPP: creative output is generated in one shot, so we provided an example. (3) AgentVerse: as in its brainstorming scenario. All methods were implemented using ChatGPT-4o mini.

Result The evaluation results are shown in Table 3. From this table, we conclude that Camel surpasses SPP and AgentVerse in ROUGE scores. This is due to Camel specifying the complex task before two agents role-play to solve individual tasks. While similar to ACT's approach, ACT involves all agents in task design, whereas Camel relies on a designated specifier. Consequently, Camel agents do not build on each other's task knowledge, leading to solutions that overlook relationships between task resolutions.

AgentVerse generated fewer words than SPP but achieved higher accuracy in three of five communities due to its collaborative decision-making, where agents engage in complex tasks before execution. However, it lacks knowledge sharing, preventing agents from building on each other's task knowledge or integrating solutions, leading to lower accuracy than ACT. Additionally, without awareness of other agents' tasks, outputs often become mere enumerations of individual solutions. SPP assigns smaller tasks to agents without knowledge sharing, further reducing accuracy.

Finally, ACT achieves the highest accuracy among the compared methods with statistical significance ($\alpha < 0.05$). This demonstrates that clarifying the complex task for all agents during the team meeting and leveraging shared knowledge enhances the accuracy of the creative output.

In addition, we evaluated ACT's ability to accumulate and reuse knowledge across complex task resolution trials. Over 100 trials per community dataset, four agents accumulated knowledge, which was carried forward and reused in subse-

Table 6: Results of ablation study.

	Rouge-L	Rouge-1	# of Words
ACT	**44.33**	**46.99**	425.35
- knowledge	44.17	46.90	425.16
- team meeting	43.90	45.05	420.81
- keywords	43.90	46.56	422.98
- task exploration	44.13	45.60	410.81
- break time	44.13	46.80	425.41
Baseline	43.06	45.39	468.85

Table 7: Trivia Creative Writing Results.

	CoT	Self-Refine	SPP	ACT
Accuracy (%)	67.1	73.9	79.9	**82.0**

quent trials. In each trial, tasks were reviewed and assigned to the agent with at least one keyword knowledge embedding highly similar to the task embedding. Importance scores were recalculated (Section 3.4.2), and each agent received 20 keyword knowledge items for solving the next complex task. The results, presented as ACT$^+$ in Table 3, show that reusing accumulated knowledge significantly improves accuracy. This underscores ACT's unique capability to accumulate and transfer knowledge across complex tasks, a feature lacking in other methods.

Human and LLM-Based Evaluation Five human experts conducted a consensus-based assessment on a randomly selected 50% of creative outputs (250 instances) from five communities. The evaluation measured how well each output aligned with the intended solutions of the four agents, considering both exact word matches and semantic consistency. We also applied a subjective evaluation using a two-point Likert scale (Likert, 1932): evaluators assigned a score of 0.5 to partially correct outputs (demonstrating semantic consistency with the correct solutions), and 1.0 to completely correct outputs (containing exact word matches). Since this reflects the task completion rate across the four agents, scores ranged from 0 to 4. Table 5 shows that ACT achieved the highest accuracy and highlights the correlation between ROUGE scores (Table 3) and human evaluation results.

We also conducted an LLM-as-a-judge evaluation using ChatGPT-4o mini to assess whether each method's creative outputs aligned with the intended solutions. As shown in Table 4, ACT achieved the highest accuracy, followed by SPP. Details are in Appendix A.4, and prompts are in Appendix C.7.

Ablation Study We performed an ablation study by shuffling 1,646 tasks across all communities and randomly assigning them to four agents, conducting 200 trials. The averaged results, presented in Table 6, show the impact of removing each function from ACT. In this table, "-" indicates the exclusion of a function; for example, "- knowledge" refers to the method that does not utilize knowledge in ACT. We also evaluated the Baseline by testing ACT without any ablation components active. Table 6 demonstrates that all functionalities contribute to ACT's performance, emphasizing the importance of team meetings for building mutual knowledge among agents while refining the complex task into specific components. It also highlights the value of abstract knowledge, as keywords facilitate creative collaboration across diverse tasks, aiding in the integration of different task solutions from a broad perspective.

Comparison of Outputs: ACT vs. SPP To assess the quality of ACT's outputs, we compared them with those generated by the baseline method, SPP. The creative outputs of ACT and SPP are presented in Tables 8 and 12, respectively, while the reference solution answers for the Tea Example are shown in Table 13. From these tables, ACT effectively integrates solution elements from multiple agents into cohesive, user-oriented proposals. In contrast, SPP largely presents isolated task outputs without synthesis. For example, ACT's business plan for the tea domain combines concrete insights from Agents A to D into unified sections such as "High-End Loose Leaf Teas" and "DIY Flavor Kits," enhancing clarity and utility. SPP's outputs, on the other hand, remain vague and less actionable. As a result, ACT captures and elaborates on more correct solutions across agents within its creative outputs, while SPP includes fewer correct elements, leading to lower accuracy. See Appendix A.3 for evaluation details and complexity analysis.

4.2 Trivia Creative Writing

We conducted the Trivia Creative Writing task to evaluate our method using the same dataset and settings with five agents as in SPP (Wang et al., 2023). This task measures accuracy by verifying whether creative outputs contain the correct answers to trivia questions, which typically have a single definitive answer. An answer matching any alias in the TriviaQA dataset was considered correct, with recall, used as a measure of accuracy, calculated as the number of correct mentions divided by the total trivia questions. We compared ACT with Chain-of-Thought (CoT) (Wei et al., 2022), Self-Refine (Madaan et al., 2023), and SPP (Wang et al., 2023). Table 7 shows that ACT outperforms existing methods, achieving higher recall by leveraging agents' knowledge sharing and cognitive synergy in solving complex tasks.

4.3 Tool Utilization Capabilities

Recent studies (Qin et al., 2023; Schick et al., 2023; Chen et al., 2024c) highlight that equipping LLMs with real-world tools significantly enhances their performance. We evaluate ACT on 10 complex Open QA tasks, each requiring the use of nine tools, including a web search API and a code interpreter, as listed in the AgentVerse GitHub repository (Chen et al., 2024c). We follow the same dataset as used in the AgentVerse paper. In our experiments, ACT decomposes a given complex task into subtasks, assigns them to agents, and enables each agent to express its opinion using tools. Agents then collaboratively integrate their insights into the creative output (Section 3.6).

For comparison, we evaluate ACT against AgentVerse and ChatGPT-4o. Since the original AgentVerse experiments were conducted with GPT-4, we use the same setting to ensure consistency with their published results. ACT, on the other hand, employs GPT-4o. We use the web interface version of ChatGPT-4o as it supports a broader range of tools, making it a strong baseline for comparison. ACT operates with four agents, iteratively calling tools to refine their opinions until a conclusion is reached or a maximum of 10 iterations is performed. Tool execution is handled via OpenAI's function calling. To assess tool effectiveness, we applied the AgentVerse evaluation criteria to determine whether ACT met them.

Our evaluation demonstrates that ACT effectively structures complex tasks within team-based interactions, enabling agents to integrate knowledge and refine solutions collaboratively. This multi-perspective approach allows ACT to generate more detailed and concrete solutions. Notably, ACT successfully solved all 10 complex tasks, while AgentVerse solved 9 and ChatGPT-4o solved 7 under the AgentVerse evaluation criteria.

Moreover, ACT enriched its outputs with more useful information through multi-perspective agent collaboration. To assess this, we introduced supplementary evaluation criteria considered valuable by the authors. These results are summarized as "further evaluation" outcomes in Tables 15–20, providing additional insights beyond the main evaluation. For example, in a complex book club planning task—detailed in Appendix B—ACT first restructured the original task through collaborative refinement among multiple agents, then elevated it into a more concrete and context-rich version of the complex task. As a result, the generated outputs included not only book summaries and access options, but also discussion prompts that were not part of the original task specification, making the outputs richer and more engaging.

ACT also mitigated hallucinations by enabling agents to cross-verify task results. By selecting titles based on the *New York Times* Best Sellers list using the search tool—and allowing multiple agents to verify the retrieved information—ACT further ensured objectivity and maintained credibility through reliance on verified sources.

5 Conclusion

This paper presented ACT, a framework that enables agents to collaboratively define complex tasks while developing structured knowledge of each other's approaches. By leveraging this knowledge, agents can effectively execute the overall task. Compared to existing methods, ACT improves solution quality through collaborative task design and execution.

We are currently exploring real-world applications of ACT, such as e-commerce business development—where agents support brainstorming, product launch planning, and marketing refinement—and end-to-end service development, where ACT manages the full lifecycle from planning to deployment. These examples are expected to demonstrate ACT's practical utility in multi-agent collaboration. To support scalability in such systems, ACT^{+} plays a key role by enabling effective knowledge reuse. Building on this foundation, we are also exploring hierarchical agent management to further reduce computational overhead in large-scale deployments. Furthermore, a key challenge lies in managing intricate sequential dependencies among agent interactions (Xu et al., 2024). We anticipate that such cases may benefit from human-AI collaboration and the introduction of intermediate checkpoints to facilitate better coordination. Finally, we plan to incorporate knowledge graphs to enhance traceability and transparency, enabling clearer reasoning and more systematic detection of hallucinated outputs.

6 Limitations

In this study, we evaluated ACT's performance across three datasets using reproducibility metrics, specifically ROUGE, recall, and human evaluation. ACT improves task accuracy by leveraging knowledge acquired during meetings, ensuring task alignment while allowing agents to focus on their assigned tasks.

LLMs can generate outputs that appear creative; however, whether they possess true creativity remains an open question (Chakrabarty et al., 2024). Rather than producing genuinely novel ideas, LLMs primarily reconstruct learned patterns based on existing knowledge. Therefore, assessing creativity requires evaluation beyond reproducibility metrics. Our study focuses on reproducibility rather than creativity, aiming to reliably and consistently execute tasks through agent collaboration.

Future research should extend ACT's collaborative framework to better capture and integrate human intent. By ensuring reproducibility while effectively incorporating human input, AI can serve as a tool that supports and enhances human creativity. Exploring how human-AI collaboration can facilitate creative processes remains a crucial challenge for future research.

References

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proc. CHI'24*, pages 1–34.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje Karlsson, Jie Fu, and Yemin Shi. 2024a. Autoagents: A framework for automatic agent generation. In *Proc. IJCAI'24*, pages 22–30.

Pei Chen, Shuai Zhang, and Boran Han. 2024b. Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Proc. NAACL-HLT'24 (Findings)*, pages 1720–1738. Association for Computational Linguistics.

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. 2024c. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *Proc. ICLR'24*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proc. ICML'24*.

Amy Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2):350–383.

Gemini-team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, and et al. 2023. Gemini: A family of highly capable multimodal models. volume cs.CL, 2312.11805.

J. R Hackman. 2002. *Leading teams: Setting the stage for great performances*. Harvard Business School Press, Boston.

Matthew Henderson, Pawel Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrksic, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. *CoRR*, abs/1904.06472.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proc. ICLR'22*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551.

Philip Kotler, Kevin Lane Keller, Mairead Brady, Malcolm Goodman, and Torben Hansen. 2016. *Marketing Management*, 3. edition. Pearson Education Ltd.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. CAMEL: communicative agents for "mind" exploration of large language model society. In *Proc. NeurIPS'23*.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana T. Hughes, Charles Lewis, and Katia P. Sycara. 2023b. Theory of mind for multi-agent collaboration via large language models. In *Proc. EMNLP'23*, pages 180–192.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023c. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *CoRR*, abs/2310.06500.

R.A. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):5–55.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

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. In *Proc. NeurIPS'23*.

Meta. 2024. Meta llama 3, https://llama.meta.com/llama3/. Accessed: June 1, 2024.

Makoto Nakatsuji and Sohei Okui. 2020. Conclusion-supplement answer generation for non-factoid questions. In *Proc. AAAI'20*, pages 8520–8527.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Alexander Osterwalder and Yves Pigneur. 2010. *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*. Wiley.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proc. UIST'23*, pages 2:1–2:22. ACM.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. 2024. Autoact: Automatic agent learning from scratch for QA via self-planning. In *Proc. ACL'24*, pages 3003–3021.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool learning with foundation models. *CoRR*, abs/2304.08354.

Eduardo Salas, Nancy J. Cooke, and Michael A. Rosen. 2008. On teams, teamwork, and team performance: Discoveries and developments. *Human Factors*, 50(3):540–547.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *CoRR*, abs/2302.04761.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, and et.al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, cs.CL, 2307.09288.

E. Tulving. 2002. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25.

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, Yibin Liu, Jialong Wu, Shengwei Ding, Long Li, Zhiwei Huang, Xinle Deng, Teng Yu, Gangan Ma, Han Xiao, Zixin Chen, Danjun Xiang, Yunxia Wang, Yuanyuan Zhu, Yi Xiao, Jing Wang, Yiru Wang, Siran Ding, Jiayang Huang, Jiayi Xu, Yilihamu Tayier, Zhenyu Hu, Yuan Gao, Chengfeng Zheng, Yueshu Ye, Yihang Li, Lei Wan, Xinyue Jiang, Yujie Wang, Siyu Cheng, Zhule Song, Xiangru Tang, Xiaohua Xu, Ningyu Zhang, Huajun Chen, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Weaver: Foundation models for creative writing. *CoRR*, abs/2401.17268.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *CoRR*, abs/2307.05300.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. NeurIPS'22*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *CoRR*, abs/2305.14688.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Zhiruo Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Keunho Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. 2024. Theagentcompany: Benchmarking LLM agents on consequential real world tasks. *CoRR*, abs/2412.14161.

A Detailed discussion with examples

This section details the complex tasks designed by ACT and the creative outputs it produces. Specifically, we examine examples of these complex tasks in Section A.1, explore how agents accumulate knowledge and how this knowledge enhances the effectiveness of the resulting creative output in Section A.2, and compare the creative outputs of ACT and SPP in Section A.3. We next use LLM-as-a-judge in Section A.4 to assess output correctness. We then discuss the computational complexity analysis in Section A.5. We also provide a brief analysis of the impact of the importance score on knowledge selection in Section A.6. Finally, we outline ACT's mechanisms for reducing hallucination and ensuring information reliability in Section A.7.

A.1 Complex tasks designed by agents

Tables 8 and 9 present the tasks of the agents, the complex tasks designed by them, and the creative outputs generated by ACT, for the tea dataset and the design dataset, respectively. This subsection discusses the benefits of revised complex tasks using examples from the tea dataset in Table 8, while the benefits of revised complex tasks for the design dataset can similarly be found in Table 9.

Tea Dataset: A Complex Task design example From Table 8, we can understand that the text generated for the revised complex task by agents is more comprehensible than merely concatenating the "task and detail" of each agent. In Table 8, the generated text for a complex task effectively fuses four different tasks.

In detail, this text outlines a complex task focused on developing a diverse range of tea offerings tailored to different customer preferences. When integrating the four heterogeneous tasks into a single complex task, comprehensiveness and feasibility are carefully considered. This ensures that the redesigned complex task maintains coherence and practicality, even when combining initially separate elements. For instance, recurring themes such as loose leaf tea, Earl Grey, fruit- or floral-infused flavors, and tea exploration and education are identified across multiple tasks. By prioritizing these commonalities, the resulting complex task achieves both internal consistency and practical applicability.

Specifically, the task incorporates each of these themes in a structured manner. It begins with the creation of a citrus-infused black tea blend using fresh orange or lemon peels, aligning with the theme of fruit- or floral-infused flavors. It then involves researching and identifying high-end loose leaf options for orange pekoe and Earl Grey tea, corresponding to the themes of loose leaf tea and Earl Grey. Furthermore, the task includes selecting floral and fruit teas that pair well with cream or almond milk, reinforcing the emphasis on fruit- or floral-infused flavors. Lastly, it integrates tea exploration and education by educating users on proper brewing techniques for loose leaf green teas. Building on this structured approach, the text further encourages collaboration and experimentation with flavors, specifically emphasizing preferences for Earl Grey and citrus blends, which directly relate to the creation of citrus-infused black tea. It also suggests exploring "fancy" versions of teas, complementing the research on high-end loose leaf options for orange pekoe and Earl Grey. Additionally, the focus on creamy iced beverages aligns with the selection of floral and fruit teas that pair well with cream or almond milk. Finally, the text reinforces tea exploration and education by guiding new customers on brewing techniques and flavor pairings, directly supporting the integration of educational aspects in the task. Overall, it presents a comprehensive approach to enhancing tea offerings while effectively integrating the four tasks held by the agent team.

In summary, by referencing the initial tasks and the designed complex task summarized in Table 8, we can observe that the complex task maintains coherence and feasibility while incorporating these four initial tasks. This integration aids agents in understanding the context of the complex task and the relationships between their individual tasks and the overall complex task.

We can also see that the creative output in Table 8, collaboratively generated by multiple agents, effectively reflects the revised complex task. As agents grasp the overall context and identify commonalities among the four tasks, their output aligns with both the redesigned complex task and the individual agents' objectives. As a result, the final business proposal organically integrates four distinct tasks into a unified solution that addresses multiple challenges simultaneously. For example, "High-End Loose Leaf Teas" includes a selection of premium loose leaf teas, some of which pair well with milk, while others cater to those with a preference for green teas. Similarly, "Indulgent Teas" and "Tasting Kits" span multiple tasks, complementing each other. This approach moves beyond fragmented solutions, ensuring that the final business proposal effectively integrates all four task solutions in a structured and coherent manner.

A.2 Exploration of tasks and knowledge acquisition for creative output

Tea dataset example Table 10 presents an in-depth exploration of expertise, tasks, acquired knowledge, and answers for Agent C in our tea dataset.

The expertise generated during agent generation in response to Agent C's initial task seems reasonable and well-assigned, as the reader can see the initial task and its details in Table 8, and generated expertise in Table 10.

As shown in Table 10, after the team meeting, User C's current task becomes more defined from the initial task, enabling Agent C to pursue it more concretely. It does not deviate significantly from the initial task; straying too far would result in a fundamentally different problem, negatively impacting accuracy. Additionally, after the break time, Agent C's task remains unchanged. This is because, following the one-on-one meeting with the expert agent during the break time, Agent C reflected on it and determined that the challenges to be addressed had not altered.

Table 10 presents three examples of knowledge entries acquired by Agent C.

The first entry is derived from the conversation with Agent A in the meeting. The second knowledge entry comes from the dialogue with Agent D, where the keyword phrases "distinct taste" from Agent A and "unique tea experiences" from Agent D lead Agent C to generate solutions that align with the correct answers: "the best experience and value out of your tea" and "a different but enjoyable tea experience." In fact, Agent C successfully incorporates these concepts such as "appreciate the distinct flavors," "providing unique value and experiences," and "creating memorable moments" into the generated creative output (see Table 8). The third entry is acquired from the one-on-one meeting between Agent C and an expert agent during break time. This knowledge, including "unique flavors", assists the agent in generating solutions that align with the correct answer: "the differences in flavor." These concepts are also integrated into the creative output (e.g. "those seeking complex flavors," "offering seasonal flavors," and "appreciate the distinct flavors"). Through meetings, agents exchange task-solving advice while incorporating knowledge of each other's tasks. This mutual exchange fosters shared solution knowledge, enhancing both individual and collaborative task resolution. As a result, agents generate synergistic creative outputs that are both concise and highly accurate, integrating overlapping insights. Additionally, acquiring missing task-specific knowledge from expert agents further improves solution accuracy.

Design dataset example Table 11 illustrates the expertise, initial task, task explorations, knowledge acquisitions, and the answers prepared for Agent A in our design dataset.

The expertise generated during agent generation in response to Agent A's initial task seems reasonable and well-assigned, as the reader can see the initial task and its details in Table 9, and generated expertise in Table 11.

As shown in Table 11, after the team meeting and during the subsequent break time, the current task for User A becomes more defined from the initial task, enabling Agent A to pursue the task more concretely. However, it does not deviate significantly from the initial task, as straying too far from the original task would result in a fundamentally different problem to solve, negatively impacting accuracy.

Table 11 also presents three examples of knowledge entries acquired by Agent A.

First one is the knowledge acquired from the conversation with Agent C in the team meeting. The keyword phrases "decorative items" and "wall decor" acquired from Agent C lead Agent A to derive the correct solution of "decorating the space with several items." In fact, Agent A successfully incorporates this concept into the generated creative output (see Table 9).

The second knowledge entry is acquired from the conversation with Agent D in the team meeting. Agent A acquires knowledge such as "modern minimalist design" and "serene environment." Following this knowledge, Agent A poses a question to the expert agent in the next break time regarding a "reading space in a quiet environment." This inquiry leads to acquiring additional knowledge, resulting in correct solutions like "armchair," as described in the next paragraph. Furthermore, the knowledge keywords are included in the generated outputs, with some of them incorporated into Agent A's correct solutions, ultimately enhancing the accuracy of the results.

The third entry is acquired from the one-on-one meeting between Agent A and an expert agent during break time. This knowledge enables the agent to generate solutions that align with the correct answers (e.g., "reading nook" and "armchairs" in "Design and implementation of a cozy reading nook, including:") and seamlessly incorporate them into the creative output (see Table 9). Similar to the Tea example, agents exchange task-solving advice through meetings while incorporating knowledge of each other's tasks. This fosters shared solution knowledge, benefiting both individual and collaborative task resolution. As a result, agents produce concise and high-accuracy creative outputs that integrate overlapping solutions. Additionally, acquiring missing task-specific knowledge from expert agents further enhances solution accuracy.

A.3 Comparison of ACT and SPP Outputs

This subsection evaluates the outputs of ACT and SPP using examples from Table 8. The outputs of ACT and SPP are presented in Tables 8 and 12, respectively. Additionally, the solution answers extracted from the Reddit dataset for the Tea Example in Table 8 are provided in Table 13.

From Table 8, ACT leverages the knowledge-sharing and the integration of four agents' task-solving approaches to design a coherent complex task. Through this process, the agents deepen their understanding of the complex task and refine their discussions. The business proposal incorporates solution answers from Agents A, B, and D (e.g., using dried orange and lemon peels, dried peels and fresh zest for citrus-infused iced tea, Earl Grey Cream and Lavender Earl Grey for milk-based teas, and high-end loose-leaf teas like Darjeeling First Flush and Assam Golden Tips). It also reflects insights from Agent C (e.g., exploring different types of green teas such as Gyokuro and

Dragon Well). The "High-End Loose Leaf Teas" section of the proposal demonstrates how ACT integrates the answers from both Agents B and C.

Additionally, the proposal's "DIY Flavor Kits" merge Agent C's task (exploring uncommon flavors) with Agent D's solutions (e.g., dried peels and fresh zest), allowing users to experiment with custom blends. The "Tasting Kits" similarly integrate and refine tasks from Agents B and C, expanding on Agents A and B's focus on diverse tea explorations. This demonstrates that ACT goes beyond merely listing solutions—it synthesizes them into a concrete proposal that enhances user experience. Furthermore, this refinement process retains key terminology and conceptual elements from the correct solution answers (e.g., flavor exploration).

In contrast, SPP produces a business proposal for "a curated selection of premium loose-leaf teas combining citrus flavors with traditional black teas." However, its content remains a direct enumeration of individual agent tasks (e.g., 1. Citrus Black Tea Blends, 2. Creamy Tea Collection, 3. Unsweetened Loose Leaf Teas, 4. Tasting Events and Workshops) without a unifying integration process. This results in lower fidelity to the original answers than ACT. For instance, SPP's "Tasting Events and Workshops" merely describes "interactive experiences for learning about brewing techniques and flavor pairings" without offering specific solutions. This suggests a lack of structured discussion and complex task reformulation, leading to vague rather than actionable outcomes.

Moreover, SPP does not explicitly mention the specific tea varieties listed in the solution answers, such as Earl Grey Cream, Darjeeling First Flush, and Assam Golden Tips. Similarly, its "Unsweetened Loose Leaf Teas" section states that it provides "guidance on brewing techniques" but lacks the detailed insights found in the solution answers. The same issue appears in its "Creamy Tea Collection" section, where it fails to specify tea pairings with milk or cream.

Overall, SPP only partially integrates the answers, leading to a proposal that is less concrete and less useful. In contrast, ACT accurately incorporates the solution answers, refines them into concrete solutions, and organically integrates responses from different tasks into a cohesive whole.

A.4 LLM-as-a-Judge Evaluation

To further assess the correctness of creative outputs, we employed an LLM-as-a-judge method using ChatGPT-4o mini. For each complex task, we asked the model to evaluate whether the outputs generated by each method correctly addressed the intended goals of the four agents involved in the task.

The prompt asked the LLM to consider both surface-level matches (e.g., ROUGE-like string overlap) and semantic relevance. For example, even if the phrasing differed from a reference answer, a response could still be judged correct if it conveyed the appropriate meaning or function.

To ensure reliability, human experts reviewed the LLM's judgments against the ground truth answers. Table 4 presents the aggregated results, showing how many times each method generated a creative output judged as correct.

ACT showed the highest alignment with intended solutions despite operating under a zero-shot setting. Interestingly, SPP, though receiving a low ROUGE score, ranked second in the LLM judge due to producing semantically appropriate answers under a one-shot setting. This demonstrates the LLM judge's ability to detect relevant content even in the presence of lexical variation.

The prompt used for LLM-as-a-judge is detailed in Appendix C.7.

A.5 Computational Cost Analysis

This section evaluates the computational cost of ACT. To address this, we compared ACT to a simple baseline (ACT with all ablation features removed; see Table 6 in Section 4.1) using two key metrics: input/output tokens and computation time. The results are presented below. It is important to note that ACT is still an experimental research-stage implementation. As such, it includes numerous debugging processes, and its computation time and token processing efficiency have not yet been fully optimized.

Table 14 presents the results. ACT requires approximately 4.5 times longer computation time and processes about 10 times more tokens than the baseline. However, this overhead is justified for several reasons.

First, ACT significantly outperforms the baseline in terms of accuracy, even though it produces more concise outputs. It is worth noting that Rouge, a metric known to favor longer outputs, still demonstrates ACT's superior performance, making its improvements even more remarkable. Statistical significance tests further confirm the robustness of ACT's performance.

Second, despite its increased complexity, ACT maintains a computation time that remains practical for real-world applications. Additionally, the cost of processing tokens continues to decrease with advancements in model efficiency, such as ChatGPT-4o-mini, making ACT increasingly feasible. By investing in accuracy-focused research now, we can establish a foundation for future advancements that leverage decreasing computational costs, further strengthening the long-term viability of ACT.

Third, to better understand the source of token consumption, we conducted a detailed analysis of token usage across each module in ACT. The breakdown is as follows:

- 10.1% — Team Meetings
- 39.2% — Break time
- 26.5% — Production Meetings
- 24.2% — Task distribution, updates, and knowledge reflection

Our analysis revealed that the primary contributor to token increase is the *Break time* mechanism, in which agents engage in an average of six-turn discussions with the Expert Agent. This process alone accounts for approximately 40% of total token usage.

Since 'Break time' is not the most critical mechanism driving ACT's performance gains—as demonstrated in the ablation results (see Table 6) —it can be removed or simplified in future iterations if necessary. However, the core strength of ACT lies in its ability to autonomously acquire and reuse knowledge across tasks, thereby improving accuracy over time. As discussed in Section 4.1 and shown in Table 3, ACT benefits significantly from the reuse of accumulated agent knowledge in complex tasks. This distributed knowledge acquisition and reuse mechanism not only improves task resolution accuracy, but also offers a promising path toward reducing token costs in future deployments, even as agent systems scale. We believe that the long-term benefits of this strategy justify the current computational cost.

A.6 Effect of Importance Score on Knowledge Selection

We conducted a brief analysis using the Fashion dataset to assess the impact of the importance score. Removing the score led to a drop in average accuracy from 43.11 to 42.64, suggesting its utility in guiding effective knowledge use (see Table 3). Notably, ACT without the importance score still outperformed all baselines (excluding ACT^{+}), but the observed decrease highlights the score's role in supporting more effective knowledge selection. These preliminary findings suggest that the importance score may serve as a useful mechanism for enhancing knowledge selection in multi-agent systems.

A.7 Mitigating Hallucination and Ensuring Information Reliability

To mitigate hallucinations and biases in agent behaviors, ACT incorporates several mechanisms aimed at improving information accuracy, team-level validation, and corrective feedback loops:

- **Team-based Task Design (Section 3.4, Procedure (2)):** Agents collaboratively design tasks by exchanging feedback and iteratively refining their approaches. This process helps identify and correct potentially flawed assumptions early in the workflow.
- **Break time Consultations with Expert Agents (Section 3.5):** During 'Break time', agents consult with domain-specific expert agents to validate information, resolve uncertainties, and fill knowledge gaps, thereby reducing hallucination risks.
- **Issue Reporting in Production Meetings (Section 3.6):** If a proposed solution fails to meet task expectations, the team leader flags the issue, triggering a reevaluation and corrective discussion to revise or improve the approach.

These mechanisms have proven effective in practice. For example, in the *Book Club* task (see Appendix B), ACT accurately identified top-selling books by consulting verified sources. In contrast, baseline models often produced hallucinated titles or authors.

For future work, we plan to enhance traceability and transparency by integrating knowledge graphs. This will support more explicit reasoning paths, help track the provenance of information, and offer a structured means of identifying and correcting hallucinated outputs.

B Detailed results on tool utilization capabilities

This section presents a detailed evaluation of tool utilization capabilities on the OpenQA dataset (Chen et al., 2024c). We here provide two representative example outputs from ten complex tasks in the OpenQA dataset. The dataset used is the same as that in AgentVerse. We compare the results among AgentVerse, ChatGPT-4o, and ACT. The evaluation criteria are identical to those used in the AgentVerse paper, assessing performance based on the completion rate of subtasks within the complex tasks. Additionally, we introduced supplementary evaluation criteria deemed useful by the authors. These are summarized as "further evaluation" results to provide additional insights.

Complex task on book club We compare the results for the following complex task: *I want to kick off a book club with my friends. Can you tell me the top 5 bestselling books this month, gather a content summary for each, and find online platforms where we can buy or borrow them?*

The outputs from AgentVerse, ChatGPT-4o, and ACT are presented in Tables 15, 16, and 17, respectively.

ACT surpasses AgentVerse and ChatGPT-4o by leveraging complex task design, allowing all agents to grasp the task's context and facilitating multi-perspective knowledge sharing for seamless integration of tasks across multiple agents. This enables agents to assist each other in task resolution, resulting in superior outcomes. Specifically, ACT more comprehensively supports book clubs by including not only book summaries and themes but also discussion questions that foster deeper engagement. It also ensures objectivity and accuracy by selecting books based on *The New York Times* Best Sellers list, maintaining relevance and credibility. Furthermore, it enhances accessibility by offering clear guidance on borrowing books for free through OverDrive and Libby, making it easier for users to obtain the selected books.

Complex task on DIY We next compare the results for the following complex task: *I've recently taken an interest in DIY home projects. Search for beginner-friendly DIY projects that can be completed over the weekend. Also, provide a list of materials required and a step-by-step guide for each project.*

The outputs from AgentVerse, ChatGPT-4o, and ACT are presented in Tables 18, 19, and 20, respectively.

Similar to the book club task, ACT surpasses AgentVerse and ChatGPT-4o by effectively structuring the complex task, ensuring agents thoroughly understand the task's requirements and collaborate efficiently. Specifically, ACT provides a more comprehensive DIY guide by incorporating not only project ideas and required materials but also cost estimates, time commitments, troubleshooting tips, and visual aids. This enhances accessibility by allowing users to select projects based on their budget and available time while ensuring successful execution with clear, step-by-step instructions. Moreover, ACT includes a shopping guide and community engagement initiatives, fostering long-term participation and knowledge sharing among DIY enthusiasts.

C Prompt used in the study

Here, we present the prompts used in our study to promote the reproducibility of ACT and facilitate the advancement of future research on multi-agent collaboration. In the following prompts, note that "$" represents a variable, while "%" denotes a comment. The symbol ">" signifies the operation of loading the variable specified by that symbol.

C.1 Agent generation

Agents receive their assigned tasks and dynamically generate expertise knowledge that corresponds to those tasks. Table 21 presents the prompts used for knowledge generation by the agents.

C.2 Designing complex task via team meeting

As described using equation (5), agents design a complex task through the meeting by employing the function fuse. Table 22 presents the prompt corresponding to the function fuse used for designing the complex task through agent collaboration.

C.3 Task exploration by agents

As described in equation (6), agents engage in task exploration through meetings. Table 23 presents the prompt used for the agents' task exploration.

C.4 Acquisition of knowledge and episodic memory

As described in equation (7), agents acquire knowledge and episodic memory through meetings by using the function $g()$. Table 24 presents the prompt corresponding to the function $g()$ used for agents' knowledge acquisition.

C.5 Agent's opinion in production meeting

During the production meeting, as described in equation (8), agents express their opinions by considering their own tasks, the knowledge they have accumulated so far, and the team's complex task, while formulating solutions to their own initial tasks to incorporate into the final output. The prompt used for this purpose is shown in Table 25.

C.6 Generating Instructions for Calling Tools by Agents

In tool utilization, each agent generates an instruction set for selecting and executing LLM function calls based on its opinion $\mathcal{R}_i$ and initial task $\mathcal{T}_{i,0}$ using the function $\mathrm{instruct}()$, as described in Section 3.6. This function is implemented using the prompt provided in Table 26.

C.7 The prompt used for the LLM as a judge

The LLM judge, after referencing the correct answers for each task, assesses whether the creative outputs from each method include the actual solutions to the complex tasks, evaluating not only word matches by using a score like Rouge but also semantic relevance. The prompt used for this purpose is shown in Table 27.

Table 8: Examples of tasks, a complex task, and a creative output for tea dataset generated by ACT.

Agents	Agent A	Agent B	Agent C	Agent D
Tasks	Find loose leaf orange pekoe or Earl Grey tea.	Find teas that pair well with cream or almond milk for a creamy, sweet treat.	Get into unsweetened loose leaf teas and exploring uncommon flavors.	Give black tea a citrus flavor.
Details	"The user is looking for 'fancy' versions of these teas," "The user is unsure if such versions exist,"	"The user has been enjoying iced Thai tea mixed with almond milk, heavy cream, and a pinch of Splenda," "The user typically prefers straight water teas but is looking for creamy combinations," "The user has plenty of floral and fruit teas but needs recommendations for teas that work well with milk and sweetener."	"The user is new to tea and has previously only tried commercial tea bags and sweetened teas from places like McDonald's and Starbucks," "The user has started with a brand called 'Private Selection' and spent around $20-23$ on various types of loose leaf teas," "The user made a mistake by using too much tea leaf in a French press, leading to a bad experience," "The user has discovered a preference for green teas."	"The user likes Earl Grey and wants to make something similar using orange or lemon", "The user tried making an oil from orange peels but it didn't turn out well."
Complex task	Develop a diverse range of tea offerings that cater to different customer preferences while providing unique value and experiences. Start by creating a citrus-infused black tea blend using fresh orange or lemon peels. Research and identify high-end loose leaf options for orange pekoe and Earl Grey tea. Select floral and fruit teas that pair well with cream or almond milk, and educate users on proper brewing techniques for loose leaf green teas. Encourage collaboration and experimentation with flavors by focusing on preferences for Earl Grey and citrus blends. Explore 'fancy' versions of teas, enjoy creamy iced beverages, and educate customers who are new to tea about brewing techniques and flavor pairings.			
Creative output (i.e. Service proposal)	**Business Proposal for Citrus Infusion Tea Co.** **Product/Service Name:** Citrus Bliss Tea Experience **Target Persona:** Our primary audience includes Tea Enthusiasts aged 25-45 who appreciate high-quality loose leaf teas, Health-Conscious Consumers seeking natural beverages, and Social Drinkers who enjoy unique blends for gatherings. This diverse demographic values quality, flavor, and educational experiences in their tea consumption. **Executive Summary:** Citrus Bliss Tea Experience aims to revolutionize the tea experience by offering a diverse range of citrus-infused black tea blends and high-end loose leaf options. Customers will enjoy a refreshing and unique flavor profile, highlighted by our signature blend featuring a 2:1 ratio of dried orange peels to fresh juice, incorporating blood oranges and Meyer lemons. Our mission is to educate customers on brewing techniques and flavor pairings, ensuring they can fully appreciate the potential of our products. **Specific Offerings:** *Citrus Bliss Black Tea:* A signature blend featuring Earl Grey with Lemon Peel and Citrus Zest Iced Tea, designed for tea enthusiasts. *High-End Loose Leaf Teas:* Selections include Darjeeling First Flush, Assam Golden Tips, Lavender Earl Grey, Earl Grey Cream, Kamairicha, Fukamushi Sencha, Dragon Well, and Gyokuro, catering to those seeking complex flavors. *Indulgent Teas:* Offerings such as Chai Tea, Rooibos Tea, Earl Grey Tea, Jasmine Tea, Lavender Tea, Peach Fruit Tea, and Strawberry Fruit Tea that pair beautifully with cream or almond milk, creating rich, creamy experiences. *DIY Flavor Kits:* Kits that allow customers to experiment with their own blends, including dried peels and fresh zest, encouraging creativity and personalization. *Brewing Workshops:* Educational sessions focusing on tea-to-water ratios, specifically 1 teaspoon per 8 ounces, steeping times of 4-5 minutes at 200 $^{\circ}$ F, and the unique characteristics of our selected teas, enhancing customer knowledge and engagement. *Tasting Kits:* Curated kits featuring our unsweetened loose leaf teas, allowing customers to explore and appreciate the distinct flavors, including floral teas like Jasmine and fruit teas like Hibiscus. **Revenue Model:** Citrus Infusion Tea Co. will be monetized through direct sales of our tea blends and kits, a subscription service offering seasonal flavors, workshops, and partnerships with local cafes. A premium pricing strategy will be implemented to reflect the quality of our offerings while appealing to our target personas. **Other Remarks:** We are committed to sustainability by using eco-friendly packaging and sourcing organic ingredients. Our online platform will foster community engagement, allowing customers to share their experiences and connect with fellow tea lovers. The Citrus Bliss Tea Experience is not just about tea; it's about creating memorable moments and a deeper appreciation for the art of tea.			

Table 9: Examples of tasks, a complex task, and a creative output generated for design dataset by ACT.

<table>
<tr><td>Agents</td><td>Agent A</td><td>Agent B</td><td>Agent C</td><td>Agent D</td></tr>
<tr><td>Tasks</td><td>Utilize a large empty corner with a window effectively.</td><td>Set up a fun and safe playspace for kids in a newly bought house.</td><td>Find reasonably priced wall decor to tie the room together.</td><td>Decorate a first post-grad apartment on a budget.</td></tr>
<tr><td>Details</td><td>“The space is about 4 feet by 3 feet,” “The room has brand new replacement couches.”</td><td>“The playspace is about 2 feet wide, 4 feet high, and $\tilde{2}$2 feet long,” “The area has the same carpet as the living room, lights installed, and little hobbit-sized doors on each end that can’t be locked,” “Kids love the area and use it for crawling, chasing, doing laps, and playing hide and seek.”</td><td>“Likes decorating tables with little plants and candle/candle holders,” “Wants nice wall pieces to complement the room.”</td><td>“Just got my first ’adult’ job,” “The user has lived on his own and has a bunch of hand-me-down furniture from his parents’ divorce – and most, but not at all, is in decent condition. Unfortunately, the furniture is just the opposite of his style,” “The user really likes West Elm, Restoration Hardware, modern yet classic furniture,” “The user would like a minimalist apartment with lots of functionality,” “IKEA has some nice stuff, however some stuff is almost as expensive as in a furniture fair, so the user is not sure if one is better than another,” “The user would like to get most of his stuff for under 2000.″</td></tr>
<tr><td>Complex task</td><td colspan="4">The team’s complex task is to create a cozy reading nook in the empty corner with a window, transforming the space into a welcoming environment for relaxation. Additionally, the team aims to design a safe and engaging playspace layout for kids that encourages active play and exploration. This playspace will measure approximately 2 feet wide, 4 feet high, and 22 feet long, featuring cozy carpeting and ambient lighting that enhance its appeal. Alongside these tasks, the team will research and curate a selection of reasonably priced wall decor options that complement the existing room aesthetics. Furthermore, the team will compile a list of essential furniture and decor items suitable for a first post-graduate apartment, ensuring that all selections align with a modern minimalist aesthetic while staying within a $2000 budget. To achieve this, the team will focus on modern yet classic furniture styles, drawing inspiration from brands like West Elm and Restoration Hardware, while also evaluating functional pieces from IKEA that fit within budgetary constraints.</td></tr>
<tr><td>Creative output (i.e. Service proposal)</td><td colspan="4">Business Proposal for Cozy Nook Creations: Transforming Spaces into Cozy Retreats
Product/Service Name: Cozy Nook Creations
A service dedicated to transforming underutilized spaces into cozy reading nooks and engaging play areas for children, while also providing curated furniture and decor solutions for young adults.
Target Persona:
Young Adults (Ages 22-30): Recently graduated individuals starting their first job, looking to furnish their first apartment with a modern minimalist aesthetic. They value functionality, style, and affordability.
Parents of Young Children (Ages 25-40): Parents seeking to create safe, engaging, and fun play spaces for their children, while also wanting to maintain a stylish home environment.
Executive Summary:
Cozy Nook Creations aims to provide a unique experience by transforming empty corners and underutilized spaces into cozy reading nooks and engaging play areas for children. Our service will enhance the aesthetic appeal of a home while promoting a nurturing environment for both relaxation and active play. We will curate a selection of affordable, stylish furniture and decor items that align with a modern minimalist aesthetic, ensuring that young adults can create a functional and beautiful living space without exceeding a $2000 budget.
Specific Offerings:
Cozy Reading Nook Design: Personalized consultation to understand the client’s vision and space requirements.
Design and implementation of a cozy reading nook, including: “Mid-century modern armchair, small side table, decorative tripod floor lamp, tall potted plant, curated wall decor (including framed botanical prints, serene landscape artwork, and personal photo gallery wall elements), soft lighting options (e.g., string lights), and cozy throws and cushions.”
Engaging Playspace Layout: Design of a safe and engaging playspace for children, including: “high-density foam mats for safety, non-toxic vibrant wall decals featuring whimsical themes, soft play mats or rugs, age-appropriate toys and activity stations, interactive wall decor (e.g., chalkboard paint or magnetic boards), and effective storage solutions for toys and books, such as bins and shelves.”
Furniture and Decor Curation: A curated list of essential furniture and decor items for a first post-grad apartment, including: “KALLAX shelving unit and MALM bed frame from IKEA, modern coffee table from Facebook Marketplace, classic armchair from a thrift store, affordable wall decor options that complement existing room decor, recommendations for plants and candle holders to enhance the cozy atmosphere, and emphasis on DIY projects for wall art and second-hand decorative items.”
Revenue Model: Coz Nook Creations will be monetized through service fees for personalized design consultations, affiliate partnerships with furniture and decor retailers to earn commissions on curated lists, a subscription model for ongoing design advice and seasonal decor updates, and workshops on creating cozy spaces and engaging play areas.
Other Remarks: We are committed to sustainability by recommending eco-friendly furniture and decor options. We aim to create an online community platform where clients can share their cozy nook transformations and engage with others for inspiration. Additionally, we will consider offering virtual consultations to reach a wider audience and provide flexibility for clients.</td></tr>
</table>

Table 10: Examples of expertise, tasks, knowledge, and answers for Agent C in the tea dataset.

Expertise	"Tea brewing techniques," "Knowledge of loose leaf teas," "Flavor profiling," "Tea preparation methods"
Task exploration	
Initial task	Educate the user on proper brewing techniques for loose leaf green teas and recommend unique flavors to explore.
Task after team meeting	Guide the user in exploring unsweetened loose leaf teas, focusing on proper brewing techniques and flavor profiling, particularly for green teas.
Task after break time	Guide the user in exploring unsweetened loose leaf teas, focusing on proper brewing techniques and flavor profiling, particularly for green teas.
Acquired knowledge from Agent A in the team meeting	
Brief summary	Orange pekoe is a type of black tea known for its fine quality and flavor. Earl Grey is flavored with bergamot oil, offering a distinct taste.
Keywords	"black tea," "Earl Grey," "bergamot oil," "distinct taste"
Acquired knowledge from Agent D in the team meeting	
Brief summary	Citrus-infused teas combine the flavors of tea with citrus fruits, enhancing the taste and aroma. This can appeal to customers seeking unique tea experiences.
Keywords	"citrus-infused teas," "flavors of tea," "unique tea experiences," "taste and aroma"
Acquired knowledge from an expert agent in the break time	
Brief summary	Kamairicha and Fukamushi Sencha are lesser-known green teas that offer unique flavor profiles. Exploring these teas can enhance our appreciation of the diversity in tea.
Keywords	"green teas", "unique flavors", "tea diversity"
Solution answer	
Answers	"Experiment with different amounts of tea leaf and steep times to find the right balance. Generally, the more leaf you use, the shorter the steep time should be," "Re-steep the tea leaves multiple times to get ***the best experience and value out of your tea***," "Try different types of green teas from various regions to explore ***the differences in flavor***," "Consider trying lighter-oxidized oolongs, such as TieGuanYin, for ***a different but enjoyable tea experience***," "Invest in one good, all-purpose teapot and spend the rest of your budget on trying various teas to determine your preferences."

Table 11: Examples of expertise, tasks, knowledge, and answers for Agent A in the design dataset.

Expertise	"interior design," "space planning," "furniture arrangement," "lighting design"
Task exploration	
Initial task	Utilize a large empty corner with a window effectively.
Task after team meeting	Create a cozy reading nook in the empty corner with a window.
Task after break time	Design a cozy reading nook in the empty corner with a window, incorporating comfortable seating, lighting, and decor.
Acquired knowledge from Agent C in the team meeting	
Brief summary	Wall decor refers to decorative items placed on walls to enhance the aesthetic of a space. It can include art, shelves, or other decorative elements that reflect personal style.
Keywords	"wall decor," "decorative items," "aesthetic enhancement," "personal style"
Acquired knowledge from Agent D in the team meeting	
Brief summary	Modern minimalist design emphasizes simplicity and functionality, often using a limited color palette and clean lines to create a serene environment.
Keywords	"modern minimalist design," "simplicity," "functionality," "serene environment"
Acquired knowledge from an expert agent in the break time	
Brief summary	Oversized armchairs are large, comfortable chairs that provide ample seating space, ideal for creating a cozy atmosphere in a reading nook.
Keywords	"oversized armchairs," "large chairs," "ample seating," "cozy atmosphere"
Solution answer	
Answers	"Turn it into a breakfast ***nook***, a ***reading nook***, or a bar area with a bar cart," "Create a hobby space, such as a painting corner," "Decorate with artwork, a plant, a ***chair***, and some curtains," "Set up a conversation area with two complementary ***armchairs*** and a small table between them, with a plant nearby."

Table 12: Creative output by SPP for the Tea Example.

Creative output (i.e. Service proposal)	**Service Name:** Citrus & Cream Tea Co. **Target Persona:** Tea enthusiasts who enjoy exploring unique flavors, particularly those who appreciate creamy tea experiences and citrus-infused blends. This includes individuals who are new to loose leaf teas and are looking for guidance in their tea journey. **Executive Summary:** Citrus & Cream Tea Co. aims to revolutionize the tea experience by offering a curated selection of premium loose leaf teas that combine citrus flavors with traditional black teas, as well as creamy blends that pair perfectly with milk. Our goal is to provide a unique and enjoyable tea experience for both seasoned tea drinkers and newcomers alike. **Experience Provided by the Service:** Customers will embark on a sensory journey through our carefully crafted tea blends, with options for personalized recommendations based on their preferences. We will offer tasting events and educational workshops to enhance their understanding of tea and its versatility. **Specific Offerings:** *Citrus Black Tea Blends:* A selection of black teas infused with dried citrus peels and bergamot oil, reminiscent of Earl Grey but with a unique twist. *Creamy Tea Collection:* A range of teas specifically designed to pair well with cream or almond milk, including floral and dessert-inspired blends. *Unsweetened Loose Leaf Teas:* A curated selection of uncommon flavors for those looking to explore the world of unsweetened teas, with guidance on brewing techniques. *Tasting Events and Workshops:* Interactive experiences where customers can learn about tea blending, brewing methods, and flavor pairings. **Revenue Model:** Our revenue will come from direct sales of tea blends, subscription boxes for regular customers, and fees from workshops and tasting events. We can also explore partnerships with local cafes and restaurants to feature our blends. **Additional Comments:** We should consider creating an online community where customers can share their experiences, recipes, and tips for brewing the perfect cup of tea. This will enhance customer engagement and loyalty.

Table 13: Solution Answers for the Tea Example.

Agent	Solution Answers
Agent A	- Use slightly crushed or ground Sichuan pepper for a unique citrusy flavor. - Add pieces of dried orange peel, separating the peel from the bitter pith and drying it. - Add lemongrass or squeeze lemon/orange into steeped tea. - Use Bergamot zest or Kaffir lime/leech lime for a similar flavor.
Agent B	- Check out Harney & Sons for loose leaf orange pekoe or earl gray tea. - Look for loose leaf versions of Twining's Earl Grey in supermarkets. - TeaGschwendner offers Earl Grey Nr. 69, which is available in the US through their website TGTea. - DAVIDsTea offers Breakfast Blend and plain Earl Grey.
Agent C	- Earl Grey is a classic tea to pair with milk. - Genmaicha and Masala Chai are good options. - A mix of Darjeeling and Assam can be used to create a Japanese-style Royal milk tea. - Lapsang Souchong pairs well with almond milk and can be sweetened if desired. - Masala Chai or breakfast blends are also good with milk. - Lavender Earl Grey with milk and sugar makes a great London Fog latte.
Agent D	- Experiment with different amounts of tea leaf and steep times to find the right balance. Generally, the more leaf you use, the shorter the steep time should be. - Re-steep the tea leaves multiple times to get the best experience and value out of your tea. - Try different types of green teas from various regions to explore the differences in flavor. - Consider trying lighter-oxidized oolongs, such as TieGuanYin, for a different but enjoyable tea experience. - Invest in one good, all-purpose teapot and spend the rest of your budget on trying various teas to determine your preferences.

Table 14: Comparison of computational time.

Model	Time (s)	Input Tokens	Output Tokens
ACT	2,649	4,169,451	389,309
Baseline	589	390,846	41,083

Table 15: Creative output by AgentVerse (book club).

Creative output	The top 5 bestsellers this month can be bought or borrowed from several online platforms. These include OverDrive, which allows you to borrow books for free through your local library or school. Other platforms include the Amazon Kindle Store, Barnes & Noble's Nook Book Store, Fictionwise, eBooks.com, eReader.com, Google Play Book Store, Powell's Books, as well as eBook lending and borrowing services like eBookFling, Lendle, and BookLending.com.
Evaluation	(Online platforms ✓, Top 5 bestselling books ✗, Content summary ✗)
Further evaluation	(Book club concept ✗)

Table 16: Creative output by ChatGPT-4o (book club).

Creative output	Starting a book club is a wonderful way to connect with friends over shared literary interests. Here are five bestselling books from February 2025, along with summaries and information on where to purchase or borrow them. **Onyx Storm by Rebecca Yarros** The third installment in the Empyrean series, this fantasy romance continues the journey of Violet and Xaden as they navigate political alliances, confront magical adversaries, and explore uncharted territories with their dragons. Released on January 21, 2025, it has quickly become a bestseller. **James by Percival Everett** A profound exploration of identity and society, 'James' delves into the complexities of its titular character, offering readers a thought-provoking narrative that challenges perceptions. This novel has garnered critical acclaim for its depth and storytelling. **The Let Them Theory by Mel Robbins** In this transformative non-fiction work, Mel Robbins introduces the 'Let Them Theory,' a concept encouraging readers to embrace acceptance and let go of unnecessary control, leading to a more fulfilling life. It's a life-changing tool that has resonated with many. **Isola by Allegra Goodman** Set in 16th-century France, this historical fiction follows Marguerite, a noblewoman abandoned in the New World. The novel chronicles her resilience and survival in an unfamiliar land, offering a rich tapestry of history and character study. **Deep End by Ali Hazelwood** A contemporary romance that dives into the world of collegiate swimming, exploring the dynamics between ambition, competition, and love. Ali Hazelwood delivers a steamy and heartfelt story that has captivated readers. **Where to Purchase or Borrow:** - Online Retailers: These titles are available for purchase in both print and digital formats n platforms like Amazon, Barnes & Noble, and Book Depository. - Local Bookstores: Check with your local independent bookstores; many offer online ordering and curbside pickup. - Libraries: Utilize your local library's catalog to borrow physical copies. Many libraries also offer digital lending through services like OverDrive or Libby. - Subscription Services: Platforms like Audible (for audiobooks) and Kindle Unlimited may have these titles available as part of their subscription packages. Starting your book club with any of these selections promises engaging discussions and shared enjoyment. Happy reading!
Evaluation	(Online platforms ✓, Top 5 bestselling books ✗, Content summary ✓)
Further evaluation	(Book club concept ✗)

Table 17: Creative output by ACT (book club).

Creative output	**Top 5 Bestselling Books for February 2025** 1. **'Onyx Storm' by Rebecca Yarros** Summary: A gripping tale that explores personal journeys in a fantasy world, focusing on resilience and adventure. Themes: Adventure, resilience, and the fight against adversity. Awards: Recognized for its engaging narrative and strong character development. Discussion Questions: - What personal experiences resonate with the themes of the book? - How do the characters' journeys reflect broader societal issues? 2. **'The Sirens' Call' by Chris Hayes** Summary: A compelling nonfiction narrative that explores social issues and public policy, advocating for social justice. Themes: Advocacy, social justice, and the power of community. Awards: Acclaimed for its impactful messaging and relevance. Discussion Questions: - What insights can we draw from the author's perspective? - How do the themes of the book relate to our own lives? 3. **'The Crash' by Freida McFadden** Summary: An intense psychological thriller that keeps readers on the edge of their seats, exploring deception and trust. Themes: Deception, trust, and the complexities of human relationships. Awards: Gained popularity for its intricate plot twists. Discussion Questions: - What are the key takeaways from the book that we can apply in our daily lives? - How do the characters' motivations drive the plot? 4. **'Hillbilly Elegy' by JD Vance** Summary: A memoir reflecting on family, culture, and the American experience, discussing identity and struggle. Themes: Identity, struggle, and social mobility. Awards: Recognized for its candid perspective and storytelling. Discussion Questions: - How does the author's background shape his views? - What societal issues are highlighted through the author's experiences? 5. **'Atomic Habits' by James Clear** Summary: A practical guide on building good habits and breaking bad ones, focusing on personal development. Themes: Personal development, productivity, and self-improvement. Awards: A bestseller for its actionable advice and clear framework. Discussion Questions: - What strategies from the book can we implement in our lives? - How do habits shape our identity and future? This information is based on the most recent data from The New York Times Best Sellers list, found at `https://www.nytimes.com/section/books/best-sellers`. **Platforms for Purchasing and Borrowing:** Purchasing Options: - Amazon: Extensive selection and competitive pricing. - Book Depository: Free international shipping. - Barnes & Noble: Membership program offering discounts. Borrowing Options: - Public Libraries: Free access to physical and digital books. - OverDrive: User-friendly for e-books and audiobooks. - Libby: Streamlined borrowing experience. **Comparison Criteria:** Formats: Options for hardcover, paperback, e-book, and audiobook. Considerations: Return policies, membership costs, and availability for each platform. This comprehensive solution addresses the challenge of kicking off a book club with friends by providing the top 5 bestselling books, their summaries, and options for purchasing or borrowing them, while also incorporating detailed discussion questions to facilitate engaging conversations.
Evaluation Further evaluation	(Online platforms ✓, Top 5 bestselling books ✓, Content summary ✓) (Book club concept ✓)

Table 18: Creative output by AgentVerse (DIY).

Creative output	The team has successfully found a variety of beginner-friendly DIY home projects that can be completed over the weekend. These projects range from painting a room, replacing cabinet pulls, installing a wall-mounted television, building a fire pit, to more complex tasks like kitchen remodels and furniture updates. The materials required for these projects are basic and easy to source, including tools like a roller and pan for painting, a cordless drill for replacing cabinet pulls, a stud finder for installing a wall-mounted television, and concrete work tools for building a fire pit. Detailed, step-by-step guides for each project have been provided to ensure a smooth DIY experience. Happy DIY-ing!
Evaluation	(DIY projects ✓, List of materials ✓)
Further evaluation	(Time commitment ✗, Cost ✗)

Table 19: Creative output by ChatGPT-4o (DIY).

Creative output	Here are some beginner-friendly DIY home projects you can complete over a weekend, each with a list of required materials and step-by-step instructions. **Create a Mood Board** A great way to visualize design ideas and gather inspiration for future projects. Materials Needed: Corkboard or foam board, magazines, fabric swatches, paint samples, scissors, glue or push pins. Steps: 1. Gather inspiration by collecting images, colors, and textures that resonate with your desired aesthetic. 2. Arrange the collected items on the board, experimenting with layouts until satisfied. 3. Secure the items using glue or push pins. 4. Display the mood board in a space where it can inspire future projects. **Install Peel-and-Stick Wallpaper** An easy and temporary way to refresh a room with minimal effort. Materials Needed: Peel-and-stick wallpaper, measuring tape, utility knife, squeegee or smoothing tool. Steps: 1. Measure the wall to determine the dimensions of the area you wish to cover. 2. Cut the wallpaper into manageable strips, adding a few extra inches for adjustments. 3. Apply the wallpaper by peeling the backing and carefully sticking it to the wall, smoothing out bubbles with the squeegee. 4. Trim excess wallpaper at the edges using the utility knife. **Transform Thrifted Furniture** Give second-hand furniture a fresh, new look with paint or stain. Materials Needed: Second-hand furniture piece, sandpaper, paint or stain, paintbrushes, new hardware (optional). Steps: 1. Prepare the surface by sanding the furniture to remove old finish and smooth imperfections. 2. Apply paint or stain evenly, allowing it to dry thoroughly. 3. Add new hardware such as knobs or handles to enhance the updated look. **Add Texture with Beadboard Paneling** Enhance walls with classic beadboard paneling for added depth and style. Materials Needed: Beadboard panels, measuring tape, saw, construction adhesive, finishing nails, paint (optional). Steps: 1. Measure the wall to determine the height and width where the beadboard will be installed. 2. Cut the beadboard panels to fit the measured space. 3. Apply construction adhesive to the back of each panel. 4. Position the panels on the wall, ensuring proper alignment. 5. Secure the panels with finishing nails into the wall studs. 6. Paint the beadboard to match your decor if desired. **Update Cabinet Hardware** A simple yet effective way to refresh kitchen or bathroom cabinets. Materials Needed: New cabinet knobs or pulls, screwdriver, measuring tape, drill (if new holes are needed). Steps: 1. Remove old hardware using a screwdriver. 2. Measure for new hardware, ensuring alignment with existing holes; if necessary, mark and drill new ones. 3. Install new knobs or pulls, tightening the screws securely. **Create a Gallery Wall** Display art and photographs in a visually cohesive way. Materials Needed: Assorted picture frames, artwork or photographs, hammer, nails, level, measuring tape. Steps: 1. Select artwork or photographs to create a cohesive collection. 2. Plan the layout by arranging frames on the floor before hanging. 3. Mark the wall using a measuring tape and level to indicate where each frame will be placed. 4. Hang frames by hammering nails into the marked spots. **Install Floating Shelves** A practical and stylish way to add storage and display space. Materials Needed: Floating shelf kit, stud finder, level, drill, screws. Steps: 1. Locate wall studs using a stud finder to ensure secure mounting. 2. Mark the placement for the shelves and check that it is level. 3. Install brackets securely into the studs. 4. Mount shelves onto the brackets and ensure they are level. **Where to Purchase Materials:** - Home Improvement Stores: Hardware and materials can be found at stores like Home Depot, Lowe's, and Ace Hardware. - Online Retailers: Many materials are available through Amazon, Wayfair, and specialty online retailers. - Local Thrift Stores: Great for finding second-hand furniture for restoration projects. Starting your DIY project with any of these ideas will add a personal touch to your home while building new skills. Happy crafting!
Evaluation	(DIY projects ✓, List of materials ✓)
Further evaluation	(Time commitment ✗, Cost ✗)

Table 20: Creative output by ACT (DIY).

Creative output	**Beginner-Friendly DIY Projects for the Weekend** **1. Wooden Planter Box** Materials Required: 1x6 untreated pine boards, wood screws, wood glue, sandpaper, exterior wood stain (optional). Estimated Cost: $20, Time Commitment: 3-4 hours, Complexity: Basic woodworking skills. Step-by-Step Instructions: 1. Measure and cut the pine boards to the desired dimensions for the planter box. 2. Sand the edges and surfaces to ensure a smooth finish. 3. Assemble the box by attaching the sides using wood screws and wood glue for added stability. 4. Optionally, apply exterior wood stain for weather protection. 5. Allow the box to dry completely before planting. Visual Aids: Diagrams showing measurements and assembly steps. Troubleshooting Tips: Ensure accurate measurements to avoid gaps in the assembly. **2. Decorative Wall Art** Materials Required: Canvas, acrylic paints, paintbrushes, palette for mixing colors. Estimated Cost: $15, Time Commitment: 3-4 hours, Complexity: Basic painting skills. Step-by-Step Instructions: 1. Prepare your workspace with all materials at hand. 2. Sketch your design lightly on the canvas. 3. Mix colors on the palette and start painting your design. 4. Allow the paint to dry completely before displaying. Visual Aids: Step-by-step images of the painting process. Troubleshooting Tips: If the paint is too thick, add a little water to thin it out. **3. Homemade Candles** Materials Required: Wax, candle wicks, essential oils, double boiler, candle molds or jars. Estimated Cost: $25, Time Commitment: 2-3 hours, Complexity: Basic melting and pouring skills. Step-by-Step Instructions: 1. Melt the wax in a double boiler until fully liquefied. 2. Add essential oils for fragrance and stir well. 3. Prepare the molds by securing the wicks in place. 4. Pour the melted wax into the molds and allow to cool completely. 5. Trim the wick to about 1/4 inch above the wax. Visual Aids: Diagrams showing the melting and pouring process. Troubleshooting Tips: If the wax cools too quickly, reheat gently to maintain pourability. **4. Simple Bookshelf** Materials Required: 1x12 pine board, wood screws, wood glue, paint or stain. Estimated Cost: $30, Time Commitment: 4-5 hours, Complexity: Basic woodworking skills. Step-by-Step Instructions: 1. Cut the pine board to the desired length for the shelves. 2. Assemble the shelves using wood screws and glue for stability. 3. Sand the edges and surfaces for a smooth finish. 4. Paint or stain the bookshelf as desired. 5. Allow to dry before placing books on the shelves. Visual Aids: Assembly diagrams and finishing steps. Troubleshooting Tips: Ensure the screws are tightened properly to avoid wobbling. **5. Wall Art with Recycled Materials** Materials Required: Old magazines, canvas board, acrylic paint, glue. Estimated Cost: $10, Time Commitment: 2-3 hours, Complexity: Basic crafting skills. Step-by-Step Instructions: 1. Cut out images and patterns from old magazines. 2. Arrange the cutouts on the canvas board to create your design. 3. Glue the cutouts in place and allow to dry. 4. Optionally, paint around the cutouts for added effect. 5. Display your artwork once dry. Visual Aids: Images of the collage process. Troubleshooting Tips: Use a strong glue to ensure the cutouts adhere well. **6. Building a Coffee Station** Materials Required: Wood for the station (e.g., pine or plywood), screws, brackets, paint or stain, sandpaper, wood glue (optional). Estimated Cost: $40, Time Commitment: 4-5 hours, Complexity: Basic woodworking skills. Step-by-Step Instructions: 1. Design your coffee station by sketching a simple layout. 2. Measure and cut the wood pieces according to your design. 3. Sand the edges to smooth out any rough spots. 4. Assemble the station using screws and brackets. 5. Finish the surface with paint or stain. 6. Install in your desired location. Visual Aids: Diagrams of the design and assembly process. Troubleshooting Tips: Ensure all measurements are accurate before cutting. **7. Hand-Painted Wallpaper** Materials Required: Wallpaper adhesive, paint (preferably water-based), brushes, level, measuring tape. Estimated Cost: $25, Time Commitment: 3-4 hours, Complexity: Basic painting and application skills. Step-by-Step Instructions: 1. Prepare the wall by cleaning the surface. 2. Measure and cut the wallpaper to size. 3. Apply adhesive to the back of the wallpaper. 4. Position the wallpaper on the wall using a level. 5. Smooth out any bubbles with a brush. 6. Let it dry completely. Visual Aids: Step-by-step images of the application process. Troubleshooting Tips: If bubbles appear, gently lift the wallpaper and reapply. **Shopping Guide** - Local Suppliers: Home Depot, Lowe's - Online Resources: Amazon, Etsy - Price Comparisons: Check local stores for competitive pricing on materials. **Community Engagement and Feedback** - A Facebook group will be created for users to share their completed projects and experiences. - A feedback survey will be developed to gather insights on project execution and suggestions for improvement. - Monthly brainstorming sessions will be organized to discuss feedback and generate new ideas based on user experiences. - A feedback report will be compiled to summarize user feedback and resultant adjustments to ensure continuous improvement of our offerings. This comprehensive guide addresses the challenge of finding beginner-friendly DIY projects that can be completed over the weekend, incorporating all participants' suggestions and tool execution results.
Evaluation	(DIY projects ✓, List of materials ✓)
Further evaluation	(Time commitment ✓, Cost ✓)

Table 21: Prompt for knowledge generation by agent.

$name is a member of the team $team_name. This team is currently collaborating in designing a complex task.
Team Name: $team_name
Now, $name recognizes the following task:
Task: $initial_task
Subject: $subject
Details:

- $detail1
- $detail2
- $detail3
- $detail4 % Add more details as necessary

In this context, what expertise does this individual possess? Please list up to N keywords of expertise.

Table 22: Prompt for fusing tasks to design the complex task.

% Opinions from each agent are listed here, detailing how each agent's task can contribute to the complex task.
> opinion
% Knowledge entries from agents for designing the complex task.
> knowledge
% Advice from each agent is listed here, focusing on the feasibility of designing the complex task based on the opinions provided.
> advice
This method combines the tasks of the participants to create a "feasible" team task.
The format should be as follows:
Subject: This section should enumerate the subjects of each participant's task: $subject.
To present these subjects as the subjects of a unified complex task, please incorporate supplementary information between each subject, or at the beginning or end of the combined subjects. This should be done while considering the opinions of each participant from the meeting record to enhance overall comprehensibility.
Detail: This section should enumerate the details of each participant's task: $detail.
To present these details as the detail of a unified complex task, please incorporate supplementary information between each detail, or at the beginning or end of the combined details. This should be done while considering the opinions of each participant from the meeting record to enhance overall comprehensibility.
Ultimately, please integrate the fused subject and detail into a complex task description that all agents in the team can understand. This will facilitate collaboration and ensure that everyone is aligned in their efforts towards an actionable and feasible solution.

Table 23: Prompt for agent's task exploration.

$teamName
You are a member of the team $team_name, and your team currently has a mission: $complex_task. In this context, you recognize the following initial task:
Subject: $subject
Details: $details
% Load your knowledge entries to design and perform the complex task.
> knowledge
Based on the aforementioned initial task, you are currently setting the following task: $current_task
Please refine your current task to fulfill the above initial task.
In this context, the definition of "current task" is:
- "current task": Specified task to fulfill your initial task.
You are currently in a meeting and are required to improve your "current task" based on the following meeting record.
{> meeting_history}
Refer to the record, understand the issues presented regarding your "current task", and make sure to include measures to address these issues in your improvements.
The knowledge you have acquired over time is crucial. You can use this accumulated knowledge to approach the needs from multiple perspectives, enhancing your initial task by integrating insights from outside your specialty.
Remember, as an agent (you), the most important goal is to develop a solution that aligns with your initial task. Reflecting this solution in the service proposal is your top priority. While integrating valuable feedback from team members into your current task, ensure that the final solution does not deviate from your initial task.
When considering your "current task", take the following steps:
1. Consider the benefits provided if your initial task is fulfilled. Envision the type of person (persona) who would receive these benefits.
2. For that persona, think about the daily life scenes where these benefits would be useful. Focus on What, When, and How.
3. Within the scenes considered in step 2, devise a task suitable for the persona you imagined in step 1. Envision the experiences provided by your task.
As a result, you need to devise:
Specified Task (explained straightforwardly): What should you do to fulfill your initial task, providing the benefits or experiences considered in the previous steps?

Table 24: Prompt for agent's knowledge acquisition.

% First, load your knowledge entries, including the expertise developed during the "agent generation."
> knowledge
% Next, load your initial task and current task.
> initial_task
> current_task
You have been in a conversation session, {$meeting_history}, with $target_user.
Now that it's over, you are about to reflect on it. Following cognitive science principles, you first need to extract episodes from this experience and then create knowledge entries that can be generalized and reused.
Initially, here is an example of the output for this task:
{
  result: [{
  subject: Paul,
  episodes: [
  "Paul presented the new product, 'SwiftScript,' which is ideal for office use, incorporating trivia about the history of pens as relayed by Charlie.",
  ...
  ],
  knowledge: [{
    brief_summary: "A pen is a tool for writing. In addition to those that use ink, there are also graphite-based options, such as pencils. The origin of the pen can be traced back to ancient Egypt,"
    keywords: "writing tool, pen variations, pen history",
    source: Charlie,
    importance: 8
    },{
    ...
    }]
  ...
  }],
}
The steps to construct the object above are as follows:
1. **Create Episodes**
An episode requires a subject. Since this is a conversation session, the speaker serves as the subject. Gather episodes for speakers **except for yourself** from the conversation log, and enumerate them per speaker. When creating episodes, focus on two points: a) what the subject did, and b) how they behaved. Ensure that the content of each episode is limited to 40 words.
2. **Create Knowledge**
From the log entries that form the basis of episodic memories, construct brief summaries about the subjects.
{ For each episode, you must generate **at least one** brief summary. This means that if there are **three** episodic memories associated with a particular individual, then at least **three or more** brief summaries must be listed under that individual. }
The basic form of each brief summary is:
"[Subject] is [definition]. [Some supplementary explanations]"
Knowledge has a hierarchical structure.
First, as higher-level information for the brief summary, extract "keywords" from the brief summary so that related information can be retrieved later. The number of keywords should be at least two and at most four.
Next, identify the "source" of the brief summary.
Finally, assign a score of importance to each knowledge entry (numeric, 0-10). This score should be based on your 'expertise' knowledge and the potential for future reuse of this knowledge entry.
Example process:
brief_summary: "A pen is a tool for writing. In addition to those that use ink, there are also graphite-based options, such as pencils. The origin of the pen can be traced back to ancient Egypt".
keywords: "writing tool, pen variations, pen history"
source: Charlie
importance: 8
Additionally, please enumerate knowledge entries per speaker in the log.
Following the above instructions, extract episodes and knowledge from the log below:

Table 25: Prompt for generating agent's opinion in the production meeting.

% First, you need to load your knowledge entries to generate creative output.
> knowledge
You are asked to articulate "what elements you find desirable in the output." This will provide material for the leader to reach a final conclusion.
When making your statement, please include the following information:
1. You have a personally assigned "initial task", $initial_task. You also have current task, $current_task, to fulfill through this production activity. Now, you have devised the solution as follows:
{$solution_of_current_task}
2. In the discussions leading up to this meeting, we have organized the team's complex task, $complex_task, which you need to contribute to in the final output. You should consider your contribution in relation to this complex task and assert it in your statements during the meeting.
{$contribution}
3. Based on the above content, please consider how we can further develop the output service proposal. You need to include specific elements (items, products, etc.) extracted from your current solution into the final output. These elements are essential for the leader when consolidating opinions.
4. Reflect on how to incorporate each component listed in step 1 into the team's creative output as a whole, ensuring coherence and synergy.
Considering the above items, please articulate your plan for realizing the team's complex task and clearly express your opinion.

Table 26: Prompt for generating instructions for calling tools by agents.

% First, load your initial task and current task.
> initial_task
Opinion: The agent has expressed the following opinion regarding the approach.
{Opinion}
The list of available tools is as follows:
{The list of available executable tools}
Task:
Identify the appropriate tools and define three sub-tasks whose results will help achieve "your initial task and opinion." Decompose the high-level query (i.e., "your initial task and opinion") into three smaller, manageable sub-tasks that can be executed using the provided tools.
Each sub-task must be:
- Small, specific, and executable
- Designed to ensure a smooth progression toward resolving your initial task
- Assigned to a unique tool from the provided list (no duplication)
Now, it's your turn. Based on this context, what sub-tasks should you define to support your initial task and opinion?
Your top priority is to resolve your initial task. When identifying and declaring sub-tasks, ensure they directly contribute to this goal.
You must generate a JSON response that STRICTLY follows the format below:
Example:
“‘json
{
  "subtasks": [
    { "arguments": ["DESCRIBE A SUBTASK"] },
    { "arguments": ["DESCRIBE A SUBTASK"] },
    { "arguments": ["DESCRIBE A SUBTASK"] }
  ]
}
“‘
Notes:
Each sub-task must be specific, achievable, and directly contribute to the team's production output.
You may provide fewer than three sub-tasks if appropriate, but no more than three.
Each sub-task must be independent and clearly defined.
Results acquired by tools should be measurable or verifiable.
Ensure that your output is a valid JSON object following the exact structure given above.
Do not include explanations or additional text outside of the JSON output.

Table 27: Prompt for the LLM as a judge.

Regarding the text of “Evaluation Target” below, please assess it against the “Expected Component”: How many concepts from the listed texts under “Expected Component” are included (scoring from 0 to 4)? In this evaluation, determine how many concepts represented by the texts listed under “Expected Component” are present in the “Evaluation Target.” If none of the concepts are included, please assign a score of 0. If all concepts from the “Expected Component” are incorporated, please assign a score of 4 (the total number of expected components). Evaluation Target: $target_sentence Expected Component (4 items): % The variable $answers lists the correct responses (i.e., four answer lists, each containing several answer candidates, since we used a non-factoid answer dataset) prepared for the initial tasks of the agents. $answers