

# FlowBench: Revisiting and Benchmarking Workflow-Guided Planning for LLM-based Agents

Ruixuan Xiao<sup>1\*</sup>, Wentao Ma<sup>2\*</sup>, Ke Wang<sup>2</sup>, Yuchuan Wu<sup>2</sup>  
Junbo Zhao<sup>1</sup>, Haobo Wang<sup>1†</sup>, Fei Huang<sup>2</sup>, Yongbin Li<sup>2†</sup>

<sup>1</sup>Zhejiang University <sup>2</sup>Alibaba Group

{xiaoruixuan, j.zhao, wanghaobo}@zju.edu.cn

{mawentao.mwt, wk258730, shengxiu.wyc, f.huang, shuide.lyb}@alibaba-inc.com

## Abstract

LLM-based agents have emerged as promising tools, which are crafted to fulfill complex tasks by iterative planning and action. However, these agents are susceptible to undesired planning hallucinations when lacking specific knowledge for expertise-intensive tasks. To address this, preliminary attempts are made to enhance planning reliability by incorporating external workflow-related knowledge. Despite the promise, such infused knowledge is mostly disorganized and diverse in formats, lacking rigorous formalization and comprehensive comparisons. Motivated by this, we formalize different formats of workflow knowledge and present **FlowBench**, the first benchmark for workflow-guided planning. FlowBench covers 51 different scenarios from 6 domains, with knowledge presented in diverse formats. To assess different LLMs on FlowBench, we design a multi-tiered evaluation framework. We evaluate the efficacy of workflow knowledge across multiple formats, and the results indicate that current LLM agents need considerable improvements for satisfactory planning. We hope that our challenging benchmark can pave the way for future agent planning research. Our benchmark is available at Github<sup>1</sup>.

## 1 Introduction

The impressive advances of large language models (LLMs) (Touvron et al., 2023; Zeng et al., 2023; OpenAI, 2023a) have spurred the evolution of LLM-driven agents (Wang et al., 2024; Hong et al., 2023) for complex task solving across diverse domains (Qian et al., 2023; Hong et al., 2023; Zhou et al., 2023). While recent works have made strides in advancing tool utilization abilities of agents through prompt construction (Yao et al., 2023b,a)

\* Equal contribution.

† Corresponding authors.

<sup>1</sup> <https://github.com/AlibabaResearch/DAMO-ConvAI>, and <https://github.com/Justherozen/FlowBench>

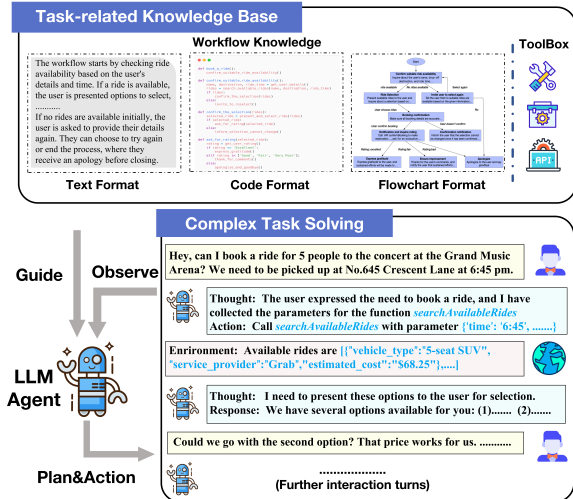


Figure 1: The procedure of workflow-guided agent planning. The agent is provided with workflow knowledge in various formats and prompted to plan the next action.

and multi-agent collaboration (Hong et al., 2023; Chen et al., 2023; Qian et al., 2023), the restricted scope of LLMs’ intrinsic parametric knowledge can lead to the undesired phenomenon of planning hallucinations (Zhu et al., 2024) – LLMs can exhibit uncontrollable actions that conflict with task knowledge and such adverse effects obstruct their practical application in knowledge-intensive tasks.

On the other hand, mitigating hallucinations by leveraging external knowledge is an ongoing key topic for LLM research (Gao et al., 2023; Baek et al., 2024; Shi et al., 2023; Baek et al., 2024). As for LLM-agent planning, some preliminary efforts have been made to enhance the planning quality of agents by incorporating workflow-related knowledge (Jiang et al., 2024; Chhikara et al., 2023; Ye et al., 2023). For example, KnowAgent (Zhu et al., 2024) employs the explicit action knowledge summarized in natural language format to regulate the planning trajectory. ProAgent (Ye et al., 2023) enhances the efficiency of complex tasks through control flows described in Python code. Despite

the promise, these nascent works commonly shape workflow-related knowledge in miscellaneous and unmethodical formats, lacking in-depth exploration and rigorous benchmarking comparisons concerning the role of workflows. Hence, it still remains underexplored how to formalize, utilize and evaluate such workflow-related knowledge for LLM-based agents across different real-world scenarios.

To this end, we meticulously explore different formats of workflow knowledge and establish a systematic benchmark for workflow-guided agent planning. Our study initially revisits and formalizes various embedding formats of workflow knowledge, including natural language, symbolic code, and flowchart schema, as depicted in Figure 1. We dub this ‘FlowAgent’ and propose its formal notations. To perform a comparative evaluation, we present **FlowBench**, the first comprehensive benchmark for workflow-guided agent planning. FlowBench covers an extensive taxonomy (6 domains, 22 roles, 51 scenarios) and different knowledge formats (text, code, flowchart) to synchronize with real-world applications. The benchmark data is constructed through a three-phase pipeline of task collection, workflow organization, and session generation. FlowBench features numerous distinct characteristics, such as coverage, difficulty, expert-level annotation, and support for multi-round user-agent interaction, as summarized in Table 1 .

To conduct a reliable assessment, we design a holistic evaluation framework that comprises two distinct granularities: (i)-Static turn-level evaluation that focuses on single-step planning. (ii)-Dynamical session-level evaluation that simulates sequential planning. Based on such evaluation mechanisms, we evaluate LLMs across varying levels of capability and embedded with different formats of workflow knowledge. We observe that even the best-performing LLM, GPT-4o (Hel, 2024), struggles to deliver satisfactory performance in certain tasks (43.2% and 40.9% success rate). By contrasting various workflow formats, we discover that flowcharts struck the best trade-off among performance, adaptability, and user-friendliness. All these results illuminate the profound challenge that FlowBench offers to the present LLM-agent designs and steer the focus of future exploration.

Our contributions are summarized as follows,

- As far as we know, we are the first to revisit and formalize the definition of different represented workflow knowledge systematically.

- We construct FlowBench, the first comprehensive benchmark for workflow-guided agent planning, which covers 51 scenarios across 6 domains, with different workflow formats.
- We present a holistic evaluation framework to achieve reliable comparisons. We provide a comparative analysis of various LLMs with different workflow formats, indicating promising directions for agent planning.

## 2 Related Work

### 2.1 LLM-Driven Agents

The evolving abilities of LLMs have fostered explorations for LLM-driven agents (Wang et al., 2024; Hong et al., 2023; Chen et al., 2023), which are designed for complex task solving and have shown great potential in various fields, such as software development (Qian et al., 2023; Hong et al., 2023; Chen et al., 2023), web navigation (Deng et al., 2023; Zhou et al., 2023), and healthcare support (Yang et al., 2024b; Zhang et al., 2023). Early approaches strive to improve the consecutive reasoning ability of LLMs, where LLMs are prompted to engage in progressive and comprehensive thinking, such as Chain of Thought (CoT) (Wei et al., 2022), ReAct (Yao et al., 2023b), and Tree of Thought (Yao et al., 2023a). The latter strand of studies focuses on enhancing the tool utilization (Schick et al., 2023; Qin et al., 2023; Patil et al., 2023) and multi-agent collaboration (Hong et al., 2023; Chen et al., 2023; Qian et al., 2023) of LLMs.

### 2.2 Knowledge-Augmented Agents

Knowledge-augmented LLMs (Gao et al., 2023; Baek et al., 2024; Shi et al., 2023; Baek et al., 2024) have been a prevailing topic, as leveraging external knowledge can effectively improve response factuality (Agrawal et al., 2023; Liu et al., 2024, 2023a). As for LLM-based agents, task-related knowledge and expertise are also important for making correct plans and decisions. To achieve this, nascent attempts (Jiang et al., 2024; Chhikara et al., 2023) have been made to enhance the reliability of planning in designated tasks by incorporating workflow-related knowledge, such as action rules described in natural language (Huang et al., 2022; Zhu et al., 2024), domain knowledge represented as knowledge graphs (Xu et al., 2024; Jiang et al., 2024), and control flow expressed in code (Yang et al., 2024a; Ye et al., 2023). However, these

Benchmark	Domain	Purpose&Task	Workflow Knowledge	Multi-turn Interaction	
				Environment	User
PlanBench	Logistics, Blocksworld	Textual Script Generation	✗	✗	✗
TravelAgent	Traveling	Travel Plan Generation	✗	✗	✗
KnowAgent	QA, Text Games	Online Task Planning	Text	✓	✗
ProAgent	Robotic Process Automation	Online Task Planning	Code	✓	✗
<b>FlowBench</b>	6 Domains, 22 Roles, 51 Scenarios	Online Task Planning	Text, Code, Flowchart	✓	✓

Table 1: Comparison of FlowBench with some related studies and benchmarks, including ProAgent (Ye et al., 2023), PlanBench (Valmeekam et al., 2023), TravelAgent (Xie et al., 2024) and KnowAgent (Zhu et al., 2024).

works mostly craft workflow knowledge in diverse and sloppy representation formats, and there is a lack of formalized consensus and comprehensive benchmarking comparisons concerning the effect of infused workflow knowledge.

### 2.3 Evaluation of LLM Agents

To assess the abilities of LLMs as agents, a plethora of benchmarks have been established (Liu et al., 2023b; Ma et al., 2024; Wu et al., 2023). These works are mostly tailored for evaluating generic abilities like tool utilization (Qin et al., 2023; Chen et al., 2024), code generation (Liu et al., 2023b; Ma et al., 2024) and embodied interaction (Wu et al., 2023). Regarding workflow-integrated planning, it remains questionable how to systematically assess the agent’s comprehension and utilization ability of the integrated workflow knowledge.

## 3 Workflow Formalization

### 3.1 Task Formulation

In our research scope, we consider an LLM-driven agent  $\mathcal{M}_\theta$  tackling a specific downstream task. This agent  $\mathcal{M}_\theta$  is equipped with external task-specific knowledge base  $\mathcal{B} = \{\mathcal{K}, \mathcal{P}\}$ , which contains workflow-related knowledge  $\mathcal{K}$  and toolbox  $\mathcal{P}$ . In most existing agent studies, agents  $\mathcal{M}_\theta$  are initially provided with complete and detailed task descriptions as initial input. Nevertheless, such a non-interactive assumption is not realistic in real-world multi-round online planning. To this end, we generalize this to a more realistic setting of multi-turn interactions, allowing users to incrementally refine and modify their requirements through ongoing conversations with the agent.

Formally, assuming that at the iteration step  $i$  during task-solving, the agent  $\mathcal{M}_\theta$  resides in state  $s_i$ . Upon receiving feedback  $e_i$  from the external environment and conversational input  $u_i$  from

the user, the agent is expected to harness both its built-in parametric commonsense  $\theta$  and external workflow knowledge base  $\mathcal{B}$  to map out an action plan  $a_{i+1}$  and provide a conversational reply  $r_{i+1}$  to the user. With this action, the agent shifts to the subsequent state  $s_{i+1}$ ,

$$\{a_{i+1}, s_{i+1}, r_{i+1}\} \leftarrow \mathcal{M}_\theta(\mathcal{H}_i, \mathcal{B}) \quad (1)$$

In this formula,  $\mathcal{H}_i$  refers to the interaction history encapsulating all the entries  $\{u_t, e_t, s_t, a_t, r_t\}_{t=0}^i$  from iteration 0 to  $i$ , which includes both current input  $\{u_i, e_i\}$  from user/environment and the historical information  $\{s_i, a_i, r_i\}$  from all past steps, as shown in Figure 1. After  $T$  turns of such sequential planning, we expect the agent  $\mathcal{M}_\theta$  to successfully accomplish the user’s tasks by formulating a correct action plan trajectory  $\{a_0, a_1, \dots, a_T\}$  and a satisfactory response  $\{r_0, r_1, \dots, r_T\}$ .

### 3.2 Revisiting Different Workflow Formats

Workflow-related knowledge is a set of representations of pipeline-related facts, which refers to the understanding of how processes or tasks are structured and executed within a specific context, such as a business or project environment. The content of different workflow knowledge primarily falls into the following categories: (i)-**Operation process related**, which describes the required steps and their sequential order to complete tasks. (ii)-**Condition/rule related**, which delineates the actions to be taken when certain conditions or parameters are met. (iii)-**Tool/data related**, which contains the utilization techniques for different tools and the mechanisms of data processing.

As mentioned earlier, workflow-related knowledge can be embedded in a broad spectrum of formats (Zhu et al., 2024; Ye et al., 2023). Thus, we begin by revisiting different workflow manifestations. Based on the level of abstraction, we primarily analyze three distinct representation forms:

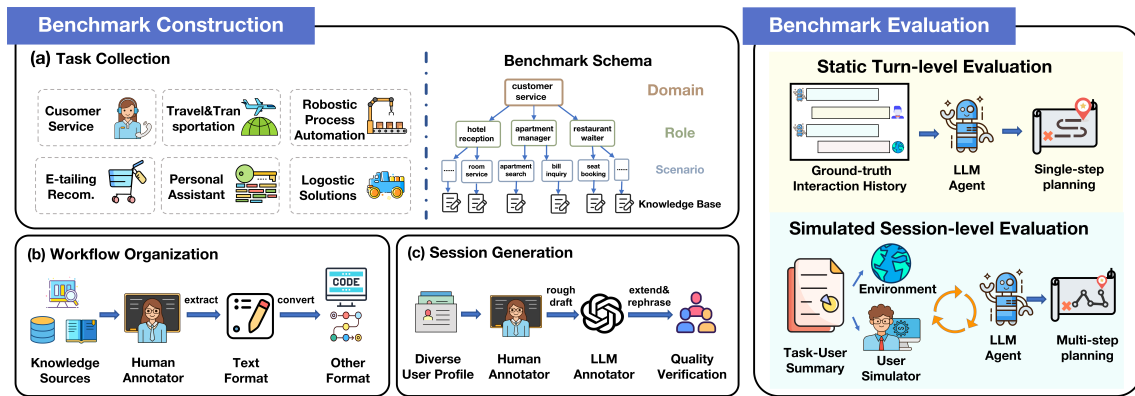


Figure 2: Overview of FlowBench. Our benchmark schema is structured in a top-down multi-level hierarchy (domain - role - scenario - knowledge). The benchmark construction process on the left contains three phases (a,b,c). The evaluation framework on the right encapsulates static turn-level and simulated session-level assessment.

- **Text** format that is conveyed through vanilla natural language documentation.
- **Code** format that follows programming language standards, e.g., Python pseudo-code.
- **Flowchart** format that is expressed in a low-code visual-programming diagram syntax, e.g., Markdown Mermaid notation.

These three forms of workflow knowledge are depicted in Figure 1 (see Appendix D for more examples). Their respective fortes and flaws can be broadly outlined with the following two criteria,

- **Efficacy and Precision:** Considering knowledge expression efficiency, text format is flexible in expressing complex concepts, but it can lead to excessive token consumption and undesired semantic ambiguity. Conversely, codes and flowcharts leverage structured symbols to enhance their precision and efficacy.
- **User friendliness:** User-friendliness is primarily contingent upon how facilitative it is for users to edit and comprehend. Although text format allows for easy editing, it can be cumbersome to pinpoint specific knowledge within lengthy documents. For code, the demand for programming expertise to edit it also presents a challenge. By contrast, the flowchart presents a more user-friendly alternative, allowing users with no coding experience to edit easily and providing visualization interfaces to improve their understanding.

Overall, it can be concluded that natural language is more expressive but less efficient and often

ambiguous, while programming code is more structured, concise, and precise, though less intuitive for user comprehension and editing. Flowcharts create a bridge, integrating the powerful expressiveness of natural language with the high efficiency of symbolic code, and providing highly user-friendly options for visualization and editing.

## 4 FlowBench

The overview of its construction process and evaluation framework is displayed in Figure 2. In what follows, we will elaborate on this in detail.

### 4.1 Hierarchy of Benchmark Schema

FlowBench consists of a collection of downstream **domains** (distribution shown in Section 4.2). Each domain incorporates a range of agent **roles**. Each of these roles is further delineated by a suite of specific **scenarios** (i.e., fundamental task). Each scenario is related to the corresponding task-related **knowledge base**, which encompasses task background, workflow-related knowledge, and a collection of tools to be employed.

A comprehensive top-down example is depicted in the left part of Figure 2. In the domain of customer services, several agent roles are configured, including restaurant waiter, hotel reception, and apartment manager. The role of apartment manager is then in charge of two scenarios: apartment search and bill inquiry. Each of these scenarios is then related to its task-related knowledge base.

### 4.2 Benchmark Construction

The construction pipeline of FlowBench is structured into three phases: scenario collection, workflow organization, and session generation.

Domain	Role	Scenario	Session	Turn
Customer service	4	12	114	1167
Personal assistant	3	7	92	821
E-tailing Recommendation	2	5	32	330
Travel&Transportation	4	9	135	1421
Logistics solutions	3	6	61	521
Robotic process automation	6	12	102	1053
Overall	22	51	536	5313

Table 2: Domain statistics of FlowBench.

**Task Collection** For a benchmark to effectively assess workflow-guided planning ability, its diversity and wide coverage are essential. To achieve this, we initially draw inspiration from the collected tasks in existing works (Mosig et al., 2020) and further conduct extensive extensions. Our task collection is targeted towards both personal customers and business enterprises. Consequently, the roles and scenarios of FlowBench are meticulously collected from the following 6 domains:

- **Customer Service** provides advice, reception, reservation, and after-sales support.
- **Personal Assistant** is to provide personal management and solution services.
- **E-tailing Recommendation** aids in product discovery and purchase during shopping.
- **Travel&Transportation** offers support for travel and transportation arrangements.
- **Logistics Solutions** assist in the management of express delivery and logistics services.
- **Robotic Process Automation (RPA)** focuses on automating and streamlining complex work processes, especially in business contexts.

As mentioned in benchmark schema (Section 4.1), each of these domains contains several agent roles (22 in total), and each role in these domains further includes several task scenarios (51 in total). The complete scenarios and roles for each domain can be found in the Appendix and the six domains are also drawn in Figure 2.

**Workflow Organization** After collecting task scenarios, next we describe the extraction and structuring of workflow-related knowledge for each scenario. Instead of brainstorming, we consult some existing references, such as professional knowledge corpora (e.g., WikiHow (Koupae and Wang, 2018)), websites of workflow knowledge (e.g., Zapier (Zapier)), and search engine results.

Integrating information from these sources, we initially summarize the workflow-related expertise point by point into a natural language document. Then different human annotators are required to independently verify the correctness, completeness, and non-redundancy of these knowledge documents. Following this, we employ GPT-4 (OpenAI, 2023b) to convert such text format knowledge into code and flowchart formats following the specified standards and incorporate manual verification to ensure knowledge consistency. An example of workflow knowledge embedded in different formats is provided in Appendix D. After structuring the workflow knowledge, we proceed to organize the tool invocation information involved in the workflow. Adhering to the GPT-4 function calling format, we outline the description for each tool call, the parameters for inputs/outputs, and the corresponding information for each parameter.

**Interactive Session Generation** With the organized workflow-related knowledge at hand, our subsequent endeavor is to synthesize the ground-truth user-agent interactive sessions for every scenario. Such generated sessions will serve as the prerequisite for our evaluation framework, whose details are discussed later in Section 4.3. Our principle is to enhance the diversity and authenticity of generated sessions. To achieve this, we prompt GPT-4 to generate diverse user profiles under each scenario to enhance diversity, which includes *user background*, *user targets*, *response tones* (i.e., response style). For data authenticity, we further intentionally incorporate some out-of-scenario items, such as casual chit-chat and irrelevant off-scope intents, into the user profiles, mimicking real-world scenarios.

Given generated user profiles, a collaborative annotation strategy is employed, which assigns a cooperative pair of a human annotator and an LLM annotator (GPT-4) for both the user and agent side. Specifically, for each interaction turn, on the user side, the human annotator crafts the subsequent user intent and dialogue draft based on task context and interaction history, which is then verified and rephrased by the LLM annotator to match the required response tone. As for the agent side, the human annotator leverages the workflow knowledge to formulate the next action plan and craft an initial response draft, which is subsequently reviewed and embellished by the LLM annotator. This collaborative generation process continues iteratively until the entire session is fully annotated.

Model	Format	SINGLE-SCENARIO					CROSS-SCENARIO				
		Tool Invocation			Parameter	Response	Tool Invocation			Parameter	Response
		P	R	F1	F1	Score	P	R	F1	F1	Score
GPT-4o	None	73.8	60.3	66.3	76.5	7.80	60.8	51.2	55.4	70.8	7.85
	Text	75.9	63.6	69.1	77.2	8.22	66.0	54.9	59.9	71.0	<b>8.34</b>
	Code	73.7	64.9	69.0	78.3	8.30	63.2	53.9	58.2	70.7	8.23
	Flowchart	<b>83.4</b>	<b>68.9</b>	<b>75.5</b>	<b>80.9</b>	<b>8.38</b>	<b>73.9</b>	<b>56.0</b>	<b>63.7</b>	<b>72.2</b>	8.29
GPT-4-Turbo	None	68.3	52.1	59.1	71.7	7.55	53.9	45.6	49.2	63.9	7.77
	Text	70.2	56.8	62.9	74.5	7.94	55.3	47.2	51.1	64.3	8.05
	Code	73.7	61.4	67.0	78.8	7.98	53.7	49.4	51.4	65.1	8.03
	Flowchart	<b>78.9</b>	<b>69.3</b>	<b>73.6</b>	<b>79.2</b>	<b>8.06</b>	<b>67.2</b>	<b>58.1</b>	<b>62.3</b>	<b>71.5</b>	<b>8.10</b>
GPT-3.5-Turbo	None	58.5	51.9	55.0	67.9	7.01	52.5	42.7	47.2	66.2	7.19
	Text	63.4	56.8	59.8	72.5	7.28	53.5	48.9	51.0	69.0	7.39
	Code	60.7	55.0	57.9	69.2	7.33	53.1	47.7	50.2	70.2	<b>7.33</b>
	Flowchart	<b>71.6</b>	<b>59.6</b>	<b>65.4</b>	<b>76.3</b>	<b>7.39</b>	<b>60.9</b>	<b>51.0</b>	<b>55.6</b>	<b>70.8</b>	7.30

Table 3: Performance comparisons of different LLMs equipped with different formats of workflow knowledge under **static turn-level** evaluation. **Bold** entries indicate superior results.

**Data Verification** To ensure data quality, we incorporate human verification at each stage during benchmark construction. Three human annotators participate in the quality verification process. Each submission from an annotator will be subject to cross-verification by the remaining two annotators.

**Benchmark Statistics** The statistics of FlowBench are shown in Table 2, and the comparisons with some related benchmarks are in Table 1. The construction cost is in Appendix A.3.

### 4.3 Evaluation Framework

After benchmark construction, we propose a holistic and multi-faceted evaluation framework in this section. We categorize our evaluation scenarios according to task awareness, and present two distinct facets of static and simulated assessments.

#### 4.3.1 Task Awareness

As previously mentioned in Section 4.1, in our dataset schema, each agent role is paired with several respective task scenarios. Therefore, we divide our evaluation scenarios into single-scenario and cross-scenario evaluations based on whether the specified task scenario is known a priori. (i)-**Single-scenario** evaluation assumes a pre-determined task scenario. An agent role is provided with the workflow knowledge of that scenario and needs to navigate, plan, and execute actions within this single task scenario. (ii)-**Cross-scenario** evaluation assumes that the specific scenario is unknown. The agent role is equipped with a versatile set of workflow knowledge covering all scenarios within the role scope. The agent needs to flexibly plan and

switch between different scenario.

#### 4.3.2 Evaluation Protocols

As mentioned in Section 3.1, our setting of multi-turn user-agent interactions generalizes the traditional agent paradigm. Consequently, the conventional evaluation protocols are incompatible with our benchmark. To solve this, we design a holistic evaluation protocol at both turn and session level, as depicted in the right of Figure 2.

**Static Turn-level Evaluation.** Static turn-level evaluation operates based on the ground-truth sessions generated in Section 4.2. Specifically, given a sampled ground-truth session  $d$ , for each turn  $i$  in  $d$ , the agent is provided with the ground-truth interaction history of previous  $(i - 1)$  turns from  $d$ , and then prompted to plan the current-turn action. Next, we compare the predicted action  $\tilde{a}_i$  and response  $\tilde{r}_i$  with ground truth  $a_i, r_i$  for evaluation.

**Simulated Session-level Evaluation.** To more comprehensively assess planning capabilities in real-world scenarios where actions are sequentially planned and carried out, a simulated session-level evaluation framework is proposed.

Specifically, we construct a user simulator based on GPT-4. In order to ensure that the behavior of this simulated user aligns with humans, we also rely on the ground-truth sessions generated in Section 4.2. We distill a task-user summary from each of these sessions via GPT-4, which concludes information of *task background*, *user goals*, *tool invocation information*. Notably, this task-user summary is different from the generated user profile in Section 4.2. The *user goals* summarizes all the

Model	Format	SINGLE-SCENARIO					CROSS-SCENARIO				
		Tool Invocation			Success	Task	Tool Invocation			Success	Task
		P	R	F1	Rate	Progress	P	R	F1	Rate	Progress
GPT-4o	None	67.8	70.5	69.0	35.1	77.5	45.1	53.9	49.0	34.2	75.7
	Text	73.7	81.2	77.2	41.7	<b>83.4</b>	<b>50.5</b>	58.7	54.1	<b>40.9</b>	81.2
	Code	70.9	79.6	75.0	<b>43.2</b>	83.1	47.7	56.9	51.9	38.2	77.9
	Flowchart	<b>74.3</b>	<b>82.6</b>	<b>78.3</b>	42.7	82.2	50.2	<b>60.4</b>	<b>54.7</b>	40.1	<b>81.3</b>
GPT-4-Turbo	None	62.9	64.0	63.5	31.9	76.7	43.6	47.0	45.2	31.1	75.3
	Text	65.9	69.9	67.8	<b>41.5</b>	<b>83.8</b>	47.6	49.3	48.4	36.8	79.6
	Code	<b>79.3</b>	64.7	71.2	37.8	83.6	47.5	<b>52.9</b>	50.0	37.3	<b>82.0</b>
	Flowchart	70.5	<b>78.3</b>	<b>74.2</b>	40.6	82.3	<b>50.1</b>	51.2	<b>50.6</b>	<b>39.4</b>	81.1
GPT-3.5-Turbo	None	37.6	50.5	43.1	29.8	63.7	24.7	30.7	27.4	22.4	61.1
	Text	44.1	<b>55.6</b>	<b>49.1</b>	<b>37.0</b>	<b>69.7</b>	<b>25.6</b>	31.5	28.2	<b>25.2</b>	63.9
	Code	<b>44.5</b>	52.9	48.3	31.3	66.3	24.5	31.9	27.8	24.1	62.9
	Flowchart	43.6	54.3	48.4	34.9	68.6	24.9	<b>33.0</b>	<b>28.5</b>	24.7	<b>67.0</b>

Table 4: Performance comparisons of different LLMs equipped with different formats of workflow knowledge under **simulated session-level** evaluation. **Bold** entries indicate superior results.

objectives of the user, while *tool invocation information* outlines the expected parameters and values for tool usage. Based on this task-user summary, a predicted session is produced with the simulated user and the to-be-assessed agent for metric calculation. All the evaluation metrics for turn and session-level are discussed later in Section 5.1.

## 5 Experiment

### 5.1 Models and Baselines

**LLMs.** Given the difficulty of solving complex tasks in practical scenarios, we mainly focus on examining the effectiveness of advanced LLMs. Hence we conduct experiments on several advanced LLMs with varying degrees of capability from OpenAI: GPT-4o (Hel, 2024), GPT-4-Turbo (OpenAI, 2023b), GPT-3.5-Turbo (OpenAI, 2022),

**Knowledge Formats.** We explore performances with different embedded formats of workflow knowledge, including (i)-**Text** variant in natural language document formats, (ii)-**Code** variant that adheres to Python pseudo-code style, (iii)-**Flowchart** variant using low-code flowcharts following the markdown mermaid syntax. (iv)-**None** variant is further included, which only offers tool information, withholding the workflow-related knowledge. Notably text/code/flowchart are based on a consistent workflow and are merely different in formats.

**Metrics.** We evaluate turn and session-level performance based on step-wise planning accuracy and overall task completion. **Tool invocation** is evaluated for both levels by precision (P), recall (R), and F1-score. An invocation is considered cor-

rect only when both tool name and all required parameters are correctly identified. On session level, we sequentially check each ground-truth invocation to verify if it matches any simulated one. On turn level, the F1-score of **parameter collection** is also included. We further evaluate **response quality** by preference scoring (full marks:10) based on correctness, helpfulness, and humanness via GPT-4. For session level, we evaluate the **success rates** and **task progress** via GPT-4. (i)-Success rate describes the proportion of completely successful sessions. A session is deemed successful when all the goals specified in the task-user summary are achieved. (ii)-Task progress indicates the percentage of goals completed within each session. The evaluation prompts are in Appendix C.3.

**Implementation Details.** During the inference, we adopt gpt-4o-0513 version for GPT-4o, gpt-4-0125-preview version for GPT-4-Turbo, and gpt-3.5-turbo-16k-0613 version for GPT-3.5-Turbo. In automatic evaluation for response preference, success rate, and average progress, we adopt gpt-4-turbo-0125-preview version for fair comparisons. The inference prompts utilize ReAct (Yao et al., 2023b) framework consisting of thought, action, observation. More implementation details and examples are put in Appendix C.

### 5.2 Main Results

Table 3 and 4 display the performance comparisons on the turn and session level. Based on these results, the following summarizations can be observed:

(i)-When external workflow knowledge is absent and merely tool information is provided (vari-

Format	S-SCENARIO			C-SCENARIO		
	Tool	SR	TP	Tool	SR	TP
None w/o tool	53.0	28.9	72.1	31.7	25.3	71.3
None	63.5	31.9	76.7	45.2	31.1	75.3
Text	67.8	41.5	83.8	48.4	36.8	79.6
Code	71.2	37.8	83.6	50.0	37.3	<b>82.0</b>
Flowchart	74.2	40.6	82.3	50.6	39.4	81.1
Ensemble	<b>75.1</b>	<b>45.3</b>	<b>85.2</b>	<b>51.8</b>	<b>43.4</b>	81.6

Table 5: Performance for variant *None w/o tool* and *Ensemble* with GPT-4-Turbo under session-level evaluation. ‘Tool’, ‘SR’, and ‘TP’ indicate F1 score of tool invocation, success rate and task progress respectively.

ant *None*), LLMs can still rely on their intrinsic commonsense to achieve a basic understanding of workflow. However, in some expertise-intensive domains, lacking workflow knowledge can result in notable degradation, as discussed in Section 5.3.

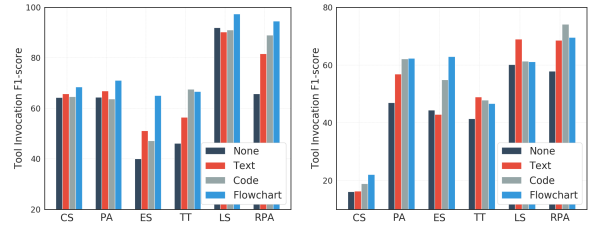
(ii)-The incorporation of workflow knowledge in different formats can significantly facilitate agent planning, achieving evident improvements at both turn level and session level. Cross-scenario transitions cause a discernible decline, yet workflow knowledge still provides a notable improvement.

(iii)-The efficacy of different knowledge formats varies in different settings. The code format is less effective on weaker LLMs, potentially because the complex symbolic expressions impede information conveyance. The text format, on the other hand, continues to perform well on different LLMs.

(iv)-Overall, the flowchart format generally produces the best performance. For static turn-level evaluations, it shows a notable advantage of 6.4%, 6.5% and 5.6% for F1 score of tool invocation on different LLMs. For session level, the flowchart demonstrates comparable performance to other formats for task completion evaluation, while it still performs the best for tool invocation. We speculate that such a lead stems from its organized and comprehensible nature, which enables LLMs to conveniently pinpoint the current state for better planning. More analyses are put to Appendix B.

### 5.3 Further Analysis

**Domain-wise Performance** We further investigate the performance distribution across domains. For experiments on GPT-4-Turbo, the domain-wise tool invocation F1-score is depicted in Figure 3. We can see that workflow knowledge plays a more significant role in those expertise-demanding domains such as RPA, E-tailing Recommendation,



(a) Single-Scenario.

(b) Cross-Scenario.

Figure 3: Domain-wise comparisons under session-level evaluation. Initials are adopted to denote domains.

and Travel, greatly compensating for the lack of inherent task-specific expertise in LLMs.

**Effect of Tool Information** As mentioned in Section 4.1, in our task-specific knowledge bases, alongside workflow-related knowledge that delineates work processes, there is also tool information describing the usage of different tools. Here, we further strip away the tool information from *None* variant and investigate this *None w/o tool* variant, from which descriptions of both tools and their parameters are completely eliminated (only tool and parameter names remain). As shown in Table 5, such a variant experiences further degradation, especially for cross-scenario settings.

**Knowledge Formats Combination.** We further explored the impact of combining different formats of workflow knowledge. As shown in Table 5, such a *Ensemble* variant is provided with text, code and flowchart formats simultaneously. It shows additional performance gains, which demonstrate that different formats of consistent knowledge can complement each other and get combined to boost the agent’s comprehension of workflows.

## 6 Conclusions

We introduce FlowBench, the first benchmark tailored for evaluating workflow-guided agent planning. We first revisit and formalize different workflow knowledge formats. Next, through extensive experiments on FlowBench, we find that flowchart formats achieve the best trade-off in terms of performance and user experience. The results further indicate that even the best-performing model, GPT-4o, fails to deliver satisfying results on challenging FlowBench. We hope that our work can provide meaningful insights to future research in the field of workflow-guided agent planning.



## Limitations

We present FlowBench, a challenging benchmark for workflow-guided agent planning. Despite its comprehensiveness, this work has certain limitations that must be acknowledged. First, our benchmark covers three representative knowledge formats from varying abstraction levels, but may not cover all potential formats, which we intend to explore in future endeavors. Moreover, we adopt GPT-4-Turbo for automatic assessment in several metrics during evaluation, which increases the evaluation cost of using our FlowBench. Our extraction process of workflow knowledge also hinges on human efforts to ensure the quality, which is quite expensive and time-consuming. Future endeavors could explore the automation of workflow-related knowledge extraction.

## Ethics Statement

In this paper, we introduce FlowBench, a highly challenging multi-domain benchmark for evaluation workflow-guided agent planning abilities. During our benchmark construction, our human annotators are instructed to filter out data potentially triggering ethical concerns, including offensive content or social prejudices. Even though the organized workflow knowledge and the ground-truth sessions have undergone filtering and verification, our simulated session-level evaluation still incorporates LLMs for simulating real-time interaction. Such API-based LLM simulators may exhibit bias and unfairness. We advise potential users to first apply bias reduction and correction methods to eliminate biased simulated sessions during evaluation, thereby enhancing fairness and ethical standards. We will release our benchmark and evaluation scripts to foster innovation and aid the development of future research.

## Acknowledgements

This work is supported by the Pioneer R&D Program of Zhejiang (No. 2024C01035). This work is also partially supported by the Fundamental Research Funds for the Central Universities (226-2024-00049) and (226-2024-00145).

## References

2024. [Hello GPT-4o](#).

Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, and Huan Liu. 2023. [Can knowledge graphs re-](#)

[duce hallucinations in llms? : A survey](#). *CoRR*, abs/2311.07914.

Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Allen Herring, and Sujay Kumar Jauhar. 2024. [Knowledge-augmented large language models for personalized contextual query suggestion](#). In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 3355–3366. ACM.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2023. [Autoagents: A framework for automatic agent generation](#). *CoRR*, abs/2309.17288.

Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, and Feng Zhao. 2024. [T-eval: Evaluating the tool utilization capability of large language models step by step](#). *Preprint*, arXiv:2312.14033.

Prateek Chhikara, Jiarui Zhang, Filip Ilievski, Jonathan Francis, and Kaixin Ma. 2023. [Knowledge-enhanced agents for interactive text games](#). In *Proceedings of the 12th Knowledge Capture Conference 2023, K-CAP 2023, Pensacola, FL, USA, December 5-7, 2023*, pages 157–165. ACM.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. [Mind2web: Towards a generalist agent for the web](#). In *NeurIPS*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *CoRR*, abs/2312.10997.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. [Metagpt: Meta programming for multi-agent collaborative framework](#). *CoRR*, abs/2308.00352.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. [Language models as zero-shot planners: Extracting actionable knowledge for embodied agents](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR.

Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, Yang Song, Chen Zhu, Hengshu Zhu, and Ji-Rong Wen. 2024. [Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph](#). *CoRR*, abs/2402.11163.

Mahnaz Koupaee and William Yang Wang. 2018. [Wikihow: A large scale text summarization dataset](#). *CoRR*, abs/1810.09305.

- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023a. [Benchmarking large language models on cmexam - A comprehensive chinese medical exam dataset](#). In *NeurIPS*.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023b. [Agentbench: Evaluating llms as agents](#). *CoRR*, abs/2308.03688.
- Xiaoze Liu, Feijie Wu, Tianyang Xu, Zhuo Chen, Yichi Zhang, Xiaoqian Wang, and Jing Gao. 2024. [Evaluating the factuality of large language models using large-scale knowledge graphs](#). *CoRR*, abs/2404.00942.
- Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. [Agentboard: An analytical evaluation board of multi-turn LLM agents](#). *CoRR*, abs/2401.13178.
- Johannes E. M. Mosig, Shikib Mehri, and Thomas Kober. 2020. [STAR: A schema-guided dialog dataset for transfer learning](#). *CoRR*, abs/2010.11853.
- OpenAI. 2022. [Chatgpt](#).
- OpenAI. 2023a. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- OpenAI. 2023b. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. [Gorilla: Large language model connected with massive apis](#). *CoRR*, abs/2305.15334.
- Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. [Communicative agents for software development](#). *CoRR*, abs/2307.07924.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. [Toolllm: Facilitating large language models to master 16000+ real-world apis](#). *CoRR*, abs/2307.16789.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *NeurIPS*.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. [REPLUG: retrieval-augmented black-box language models](#). *CoRR*, abs/2301.12652.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *CoRR*, abs/2302.13971.
- Karthik Valmeekam, Matthew Marquez, Alberto Olmo Hernandez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. [Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change](#). In *NeurIPS*.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. [A survey on large language model based autonomous agents](#). *Frontiers Comput. Sci.*, 18(6):186345.
- Jason Wei, 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 *NeurIPS*.
- Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. 2023. [Smartplay : A benchmark for llms as intelligent agents](#). *CoRR*, abs/2310.01557.
- Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. [Travelplanner: A benchmark for real-world planning with language agents](#). *CoRR*, abs/2402.01622.
- Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Kang Liu, and Jun Zhao. 2024. [Generate-on-graph: Treat LLM as both agent and KG in incomplete knowledge graph question answering](#). *CoRR*, abs/2404.14741.
- Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R. Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, Heng Ji, and Chengxiang Zhai. 2024a. [If LLM is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents](#). *CoRR*, abs/2401.00812.
- Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. [Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue](#). In *AAAI*, pages 19368–19376. AAAI Press.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of thoughts: Deliberate problem solving with large language models](#). In *NeurIPS*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. [React: Synergizing reasoning and acting in language models](#). In *ICLR*. OpenReview.net.

Yining Ye, Xin Cong, Shizuo Tian, Jiannan Cao, Hao Wang, Yujia Qin, Yaxi Lu, Heyang Yu, Huadong Wang, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2023. [Proagent: From robotic process automation to agentic process automation](#). *CoRR*, abs/2311.10751.

Zapier. [Zapier | automation makes you move forward](#).

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. [GLM-130B: an open bilingual pre-trained model](#). In *ICLR*. OpenReview.net.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. [HuatuogPT, towards taming language model to be a doctor](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. [Webarena: A realistic web environment for building autonomous agents](#). *CoRR*, abs/2307.13854.

Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Ningyu Zhang, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. [Knowagent: Knowledge-augmented planning for llm-based agents](#). *CoRR*, abs/2403.03101.

## A Additional Benchmark Details

### A.1 Entire List of Roles and Scenarios

As mentioned in Section 4.2, we collect our tasks from six separate domains, with each domain encompassing different roles and scenarios. Here we provide the entire of roles and scenarios within each domain, as shown in Table 6.

### A.2 Entire Benchmark Statistics

Here we further provide the entire benchmark statistics in Table 8, including the number of sessions and turns under the different settings of single-scenario and cross-scenario respectively.

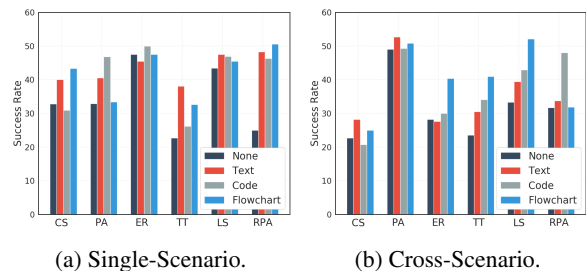


Figure 4: Domain-wise success rate under session-level evaluation. Initials are adopted to denote domains.

### A.3 Benchmark Construction Cost

We further provide the rough cost of benchmark construction. Before quality filtering, we manually constructed a total of 58 scenarios and 730 dialogues. The annotation cost is \$17.23 per scenario and \$9.44 per dialogue, amounting to a total annotation cost of \$7,892.60. Additionally, the cost for generating multi-turn dialogue requests using GPT-4-Turbo is approximately \$1,000, bringing the total cost to around \$8,892.60.

### A.4 Formatting Standards

In the process of organizing workflow knowledge, we need to convert the text-based natural language documents of workflow knowledge into code and flowchart formats. Different standard requirements are maintained for them. (i)-For code format, we adopt Python-style pseudo code, mandating that various action steps be delineated within distinct functions. The variable names and parameters must correspond to the workflow implications from text documents. For commands that are difficult to express through code, minimal commenting is permitted for clarification. (ii)-For flowchart format, we utilize Markdown Mermaid syntax for its visual operation capabilities. In the flowcharts, nodes correspond to distinct states, outlining within them the workflow directives applicable in those states, which entails rules of actions and responses. Edges represent the possible movements between nodes, with transition conditions based on user intents or environment feedback.

### A.5 Demonstration of Samples

We provide a demonstration of samples, including different knowledge formats, tool information and complete ground-truth sessions in Appendix D.

Domain	Roles	Scenarios
Customer Service	restaurant_waiter hotel_reception apartment_manager gas_equipment_service	Restaurant Search, Restaurant Booking Hotel Search, Hotel Booking, Hotel Room Service Apartment Search, Schedule Viewing, Bill Inquiry, Payment Gas Repairs, Gas Bill Payment, Gas Interruption Feedback
Personal Assistant	medical_consultant meeting_arrangement financial_assistant	Schedule a Medical Consultation, Obtain Diagnostic Results Meeting Initiation, Meeting Reschedule Currency Exchange, Withdrawal Appointments, Balance Inquiry
E-tailing Recommendation	online_shopping_support computer_store_sale	Product Search, Cart Management, Order Processing Laptop Recommendations, Computer Maintenance
Travel&Transportation	ride_service driving_service flight_inquiry travel_assistant	Ride Booking, Ride Inquiry Driving Consultation, Cancel Reservation, Modify Reservation Flight Information Search, Flight Booking Travel Guidance, Weather Inquiry
Logistics Solutions	express_support moving_service food_delivery_service	Express Delivery, Express Tracking Inquiry Moving Service Appointment, Insurance Claim Online Questionnaires, Get Food Voucher
Robotic Process Automation	invoice_management mail_administration printing_service attendance_arrangement seal_management workstation_applicant	Invoice Administration, Invoice Reimbursement Mail transmission, Mail Response Document Printing, Printing State Notification Attendance Anomaly Detection, Shift Handover Seal Request, Seal State Notification Workstation Replacement, Workstation Change Directive

Table 6: Detailed Scenarios of different domains and roles for FlowBench.

Scenario Setting	Format	Type of Failure Reasons				
		1	2	3	4	5
Single	None	52.1	4.5	6.0	35.1	2.3
	Text	52.2	3.3	2.2	40.2	2.1
	Code	57.3	1.9	2.9	35.9	2.0
	Flowchart	56.9	0.9	3.9	34.9	3.4
Cross	None	32.4	4.3	9.0	54.3	0.0
	Text	45.8	1.3	4.2	46.7	2.0
	Code	41.4	1.2	8.0	47.1	2.3
	Flowchart	49.5	0.9	7.5	40.9	1.2

Table 7: Analysis of different failure reasons for session-level evaluations with GPT-4-Turbo. This shows the percentage share (%) of different failure reasons.

## A.6 More Details of Session Generation.

In Section 4.2, we have introduced some details of interactive session generation. Here we further discuss our framework from a broad viewpoint. As shown in Figure 5, we adopt a self-talk framework driven by role-asymmetry, where both the user and agent side consists of a human-LLM collaborative pair. The user side engages in role-playing based on their profiles to present needs, and the agent side applies workflow expertise to handle processes and satisfy those needs. More details are in Section 4.2.

## B Additional Experimental Results

### B.1 Additional Performance Distribution

We further additional performance distribution across domains. For experiments on GPT-4-Turbo, the domain-wise success rate is depicted in Figure 4. It can be observed that workflow knowledge has a substantial impact in some expertise-demanding domains such as RPA, E-tailing Recommendation, and Travel&Transportation. Some domains show a noticeable performance drop in the cross-scenario setting compared to the single-scenario setting. We speculate this is due to a larger scenario gap within the domain.

### B.2 Analysis of Failure Causes

We further analyze the reasons for task failure in session-level evaluation. Based on the workflow content and observations of the agent predictions, we categorize the failure reasons into the following five categories, (i)-Type 1: Missing steps, which means the agent missed a step needed to complete the task. (ii)-Type 2: Incorrect sequence, which means the agent employs incorrect order of multiple steps. (iii)-Type 3: Incorrect transition, which includes incorrect recognition of user intent and transition conditions. (iv)-Type 4: Tool usage,

Domain	Roles	Scenarios	Sessions		Turns	
			Single-Scenario	Cross-Scenario	Single-Scenario	Cross-Scenario
Customer Service	4	12	66	48	619	547
Personal Assistant	3	7	59	33	510	311
E-tailing Recommendation	2	5	21	11	199	131
Travel&Transportation	4	9	78	57	764	657
Logistics Solutions	3	6	38	23	299	222
Robotic Process Automation	6	12	73	29	703	350
Total	22	51	335	201	3094	2219

Table 8: Entire domain statistics of FlowBench.

Task	GPT-4o	GPT-4-Turbo	GPT-3.5-Turbo
Single-Scenario	92.5	91.8	77.9
Cross-Scenario	84.0	70.7	57.3

Table 9: Comparisons of node prediction accuracy for turn-level flowchart-guided evaluation.

which means the agent fails in calling the correct tool name or in collecting parameters. (v)-Type 5: Other reasons. We conduct statistics on the reasons for failure in the samples via GPT-4-Turbo. Table 7 illustrates the proportion (%) of failure causes in each setting for GPT-4-Turbo. It can be observed that lacking workflow guidance often leads to an increase in step sequencing (type 2) and transition errors (type 3), while flowchart formats are effective in addressing sequence errors. Further, with more tools available in cross-scenario situations, there’s also an increase in tool invocation errors.

### B.3 Average Number of Turns

We further display the average number of turns for the simulated sessions during session-level evaluation in Figure 6 (each turn consisting of either an agent-user or agent-environment interaction). It can be observed that the lack of workflow results in slightly fewer rounds required (since the agent is utterly clueless about what actions to perform), whereas various formats have a negligible impact. For different LLMs, GPT-3.5-Turbo demands a larger amount of turns, with cross-scenario cases typically requiring more rounds than single-scenario as well.

### B.4 Results of Node Prediction

Due to the special graph-structured nature of flowcharts, more analysis results are possible. For flowchart-guided planning, every step involves moving to a state node. We provide node accuracy metrics for turn-level flowchart-guided plan-

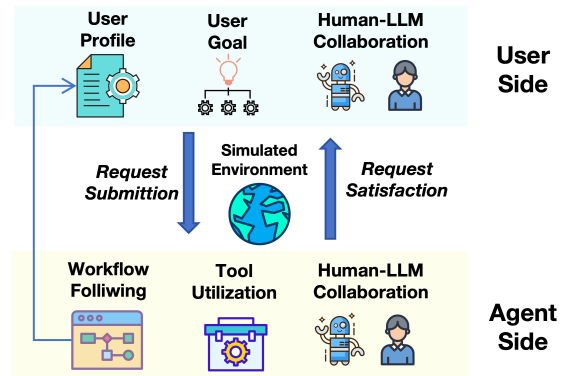


Figure 5: The session generation framework.

ning, illustrated in Table 9. It can be observed that the most capable LLM GPT-4o can achieve satisfactory performance, while there is a significant variation among LLMs of different capabilities, indicating substantial differences in their ability to comprehend structured flowcharts.

### B.5 Additional Results on Open-source LLMs

Our FlowBench is designed to provide a general and systematic evaluation framework that does not rely on specific LLM architectures (as it simply collects predictions from black-box LLMs for evaluation). Here we further provide the performance comparisons under static turn-level evaluation on the open-source LLMs Qwen2-72B-Instruct and Qwen2-7B-Instruct with different sizes to validate the extensive adaptability of our FlowBench. The results are shown in Table 10.

### B.6 Additional Alignment Verification of Human and LLMs as Judges

We further conduct an alignment verification between the judgment of humans and LLMs. To achieve this, two groups of experts are recruited to separately offer a batch of human judgments of response quality. Specifically, for both single-

Model	Format	SINGLE-SCENARIO					CROSS-SCENARIO				
		Tool Invocation			Parameter	Response	Tool Invocation			Parameter	Response
		P	R	F1	F1	Score	P	R	F1	F1	Score
Qwen2-7B	None	32.5	38.0	35.0	55.4	6.35	25.8	28.0	26.9	46.7	5.99
	Text	36.8	<b>38.5</b>	37.6	57.1	<b>6.58</b>	28.3	<b>29.7</b>	29.0	49.4	6.20
	Code	36.8	37.3	37.0	56.0	6.53	28.5	29.0	28.8	48.2	<b>6.25</b>
	Flowchart	<b>39.1</b>	<b>38.5</b>	<b>38.8</b>	<b>57.2</b>	6.56	<b>29.5</b>	<b>29.7</b>	<b>29.6</b>	<b>49.7</b>	6.22
Qwen2-72B	None	46.2	63.2	53.3	75.6	7.36	33.3	54.7	41.7	71.6	7.04
	Text	47.6	65.8	55.3	<b>78.8</b>	7.57	34.2	<b>55.2</b>	42.2	<b>72.5</b>	7.29
	Code	46.7	<b>67.5</b>	55.2	78.7	<b>7.69</b>	35.7	56.3	43.5	72.0	7.35
	Flowchart	<b>52.4</b>	65.0	<b>58.1</b>	<b>78.8</b>	7.65	<b>42.1</b>	55.1	<b>47.4</b>	72.1	<b>7.42</b>

Table 10: Performance comparisons on Qwen2-7B and Qwen2-72B equipped with different formats of workflow knowledge under **static turn-level** evaluation. **Bold** entries indicate superior results.

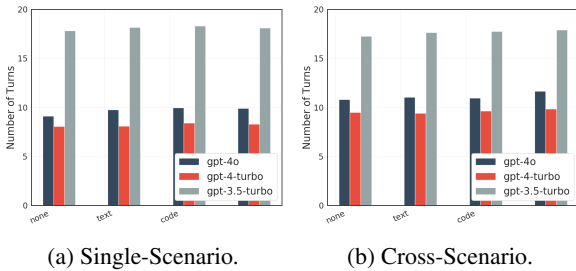


Figure 6: Average number of generated turns for simulated sessions under session-level evaluation.

scenario and cross-scenario settings, we randomly picked 50 samples and their turn-level responses on GPT-4-Turbo with 4 different knowledge formats. The human evaluators are given the same instructions on scoring criteria as the LLM evaluator. The two groups of human annotators provide independent scores for the same sampled subset. Finally, the human-model alignment verification is achieved by comparing the Pearson correlation coefficients (i)-between the human ratings of group 1 and human ratings of group 2 and (ii)-between the GPT-4 ratings and human ratings. The Pearson coefficient results are shown in Table 11. Each cell in the table is formatted as (i) human-human / (ii) human-LLM. It can be observed that the correlation coefficients between humans and LLMs are just slightly smaller than those between humans and humans (even bigger in two cases), which proves that employing LLMs for evaluation is both effective and well aligned with human judgment in our task.

## C Additional Implementation Details

### C.1 More Implementation Details

In the construction process of our benchmark, we filtered out 12% of the scenarios and 26% of the

(i)/(ii)	None	Text	Code	Flowchart
S-Scenario	0.87 / 0.81	0.88 / 0.86	0.80 / 0.76	0.82 / 0.78
C-Scenario	0.86 / 0.87	0.83 / 0.76	0.89 / 0.82	0.83 / 0.85

Table 11: Alignment Verification of human and LLMs as judges, where (i) is the Pearson coefficient between human and human, (ii) is that between human and LLM.

sessions after the initial annotation. During the ground-truth session generation, we also follow the setting of task awareness and generate sessions of both single-scenario and cross-scenario settings. In turn-level assessments, all ground-truth sessions are utilized. For simulated session-level evaluation, however, we preserve only those portions of the ground-truth sessions that are lengthy and more comprehensive, and then a user-task summary is created for simulation. For session-level assessment, the maximum number of interaction turns is restricted to 15 for single-scenario and 20 for cross-scenario. If the task is not completed within these limits, it is deemed a failure.

Moreover, we further provide additional evaluation details. During the evaluation of tool invocation, parameter matching is required. For types like numbers and enumerations, an exact match between predictions and ground truth is employed. In cases of more substantial strings where similar phrases could imply the same idea, we resort to using GPT-4-turbo for fuzzy matching, guided by precedent work. During our evaluations, All experiments are conducted three times, and the mean results are displayed.

### C.2 Prompt Design for Inference

Below we provide the inference prompt with the ReAct strategy. Notably, for both turn-level

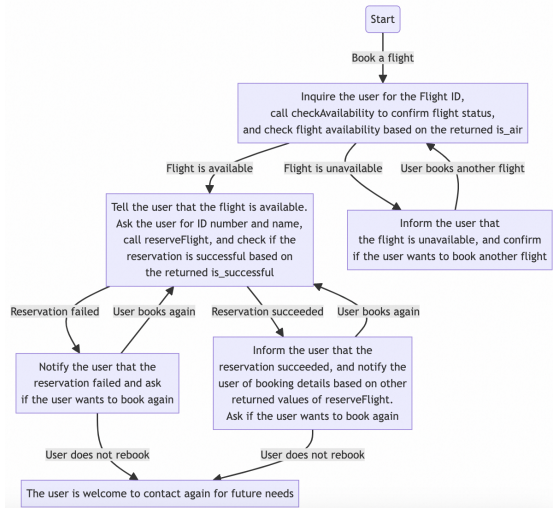


Figure 7: Visualization of flowchart format workflow knowledge for the scenario of flight booking.

and session-level evaluations, the same inference prompt is used. For the cross-scenario setting, all the task-related knowledge bases from relevant scenarios are sequentially inserted into the prompt.

```

You are a helpful assistant for the task of ....

{task background}

Specific requirements:
1. You need to act as an assistant and engage in a conversation with the user, following the business process and API information.
2. You have been provided with the flowchart information for different scenarios under a specific role.
3. You can only answer questions within the scope of the given several workflow processes. If the user asks a question beyond these scopes, please apologize and explain to the user in the response part.
4. When asking for API input parameters, ensure that the provided parameter values comply with the specified format regarding both the correctness of the format and the completeness of the content. Do not assign values arbitrarily. In instances where the parameters do not meet the format requirements, notify users to make the necessary adjustments until the requirements are satisfied.
5. When the user has multiple requests at the same time, please select one appropriate request for processing first and inform the user that other requests will be resolved subsequently. If there is unfinished business in the previous conversation, continue to provide the necessary help and guidance to assist them in completing the business process. When multiple APIs need to be called, do so in separate rounds, with a maximum of one API call output per round. When the user indicates that the business is finished or says goodbye, respond politely and end the conversation.
6. Your output format should be chosen from one of the two templates below (7.1 and 7.2):
7.1 If you need to interact with the user:
...
Thought: xxx (description of your thought process )
Response: xxx (the content you need to inquire or reply)
...
[Format Explanation]
(1) Thought includes 2 pieces of information: [Step: Analyze the current intent]: 'The current intent intent=xxx'. [Step : Decide the follow-up actions]: 'Next, I need to xxx.'
7.2 If you need to call an API (only one API call per time):
...
Thought: xxx (description of your thought process )
Action: xxx (the function name to be called, do not add the prefix "functions.")
Action Input: xxx (the parameters for the function, must be in strict JSON format)
...
[Format Explanation]
(1) 'Thought' includes the information described in sections (1) and (3) of 7.1, totaling seven [Step: xxx] pieces of information.
(2) In template 7.2, do not output 'Response', only output 'Thought', 'Action', and 'Action Input'.

```

```

8. When multiple possible intents exist, connect them with ' OR ' and clarify which situation is being inquired about; when expressing multiple intents, connect them with ' AND '.

Please adhere strictly to the defined output format. Do not output the placeholder "...." verbatim; instead, fill it with the corresponding content.

Workflow information:
{workflow information in specified formats}

Tool information:
{toolbox information}

Current time:
{current time}

```

Listing 1: Prompt for inference

### C.3 Prompt Design for Automated Evaluation.

**Response Scoring.** Below we provide the evaluation prompt of response scoring for turn-level evaluations. The full score is 10 points.

```

Please serve as an impartial judge to evaluate the response quality of the assistant. Your evaluation should be based on the following criteria:
(1) Correctness: Does the reply remain consistent with the workflow knowledge without any contradictions?
(2) Helpfulness: Has the user's request been reasonably understood and addressed, fulfilling the user's needs within the provided workflow scope?
(3) Humanness: Is the response coherent, clear, complete, and does it include human acknowledgment?
Please compare the provided response with the reference response and evaluate it based on the mentioned dimensions. Then, aggregate these assessments to assign an overall score. A perfect score is 10 points, with 9-10 points indicating high quality, nearly identical to the reference answer; 7-8 points indicating quality close to the reference answer; 6-7 points being of moderate quality; 4-5 points indicating a lower quality response; and 2-3 points for a response with significant errors.

Here is the true value response from the reference: {reference_input}

Here is the generated response from the assistant: {predicted_input}

Here is the knowledge related to the workflow: {text format workflow knowledge}

Just reply with the score, the format is as follows, Score: xxx

```

Listing 2: Prompt for response scoring

**Task Completion.** Below we provide the evaluation prompt for evaluating task completion, including the success rate and task progress.

```

You serve as an assistant responsible for assessing if a dialogue system has achieved the user's goals. You are given the provided user profile, user objectives, and the dialogue record between the system and the user, your task is to determine if the system has met all the goals of the user.

Below is the user profile, the user's objectives, including the APIs the user expects to be called with the corresponding input parameters:

User target: {user target from the task-user summary}

Below is the workflow information (mermaid) and API information of the task where the dialogue is located.

workflow_info: {text format workflow knowledge}
...

Below is the interaction content between the role of 'user' and the 'assistant' system. In the assistant's 'Thought,' the content of 'Action' and 'Action_input' indicate the API and parameters that need to be called. The content of 'function' denotes the returned results of the API calls.

simulated_session: {the predicted session}

```

Now, your task is to decide whether the dialogue has fulfilled all the user's goals previously mentioned. This encompasses whether the dialogue has completed the information inquiry and interaction needed, if the corresponding APIs have been correctly called, and whether the API call parameters were transmitted accurately.

You only need to check whether the target within the provided workflow chart and API information has been completed. If the target goes beyond the scope of the provided workflow information, there is no need to check this target. You don't need to check whether the return value of the API parameter is reasonable, you only need to check whether the parameter collection is reasonable.

Only consider whether the system (assistant) has completed the task. If the user does not make any requests in the goal, or makes requests beyond the goal, the goal is considered completed. The main criteria are the collection and invocation of API parameters. You need to focus on examining whether the intent parsing and responses are reasonable, and whether each goal is completed. If all goals are completed, it is considered successful

You also need to output the total number of user goals (if there are multiple goals), and the number of goals that have been achieved, considering goals outside the business process as completed. (The total number of goals is greater than or equal to the number of completed goals). Use a rigorous evaluation mode where each goal is achieved (or out of scope) for overall success

Your reply template should follow this format:  
Result: yes/no (overall success)  
Total number of goals: 4  
Number of accomplished goals: 3  
Reason: If the goal is completed, no reason is required. If not, please specify why you judged it to be incomplete.

Listing 3: Prompt for assessing task completion

#### C.4 Additional Annotator Information

**Annotator Information** Below we provide some basic information about the annotators.

- **Number of annotators:** Overall 17 annotators participated in the construction of our benchmark, where 5 of them were in charge of task collection and workflow organization, 9 were responsible for interactive session generation, and 3 were involved in data verification.
- **Annotation expenses:** The complete data construction process took the annotation groups about 2.5 months. The annotation expenditure is also shown in Appendix A.3.
- **Demographic information:** Among these 17 annotators, in terms of age, 10 are between 20-29 years old, 4 are between 30-39, and 3 are between 40-49. Regarding gender, 12 are female and 5 are male. In terms of educational level, 13 annotators possess a bachelor's degree, while 4 have a graduate degree. All these annotators have been provided with exhaustive annotation training.

**Annotator Instruction** We adopt a human-LLM collaborative annotation strategy for interactive session generation. In practical annotation, the human annotators are given an instructional manual and

undergo training based on already annotated examples. Below we provide a brief instruction.

```
Objective:
Based on the existing conversation history, please formulate
the user intent and response draft in accordance with the
user background and targets specified in the user profile.
Annotation steps:
(a) Please carefully read the entire content of the user
profile, focusing particularly on the user's background and
targets.
(b) Based on the existing conversation history, analyze
which targets outlined in the user profile remain incomplete
, and determine the next target to focus on.
(c) Think from the user's perspective and identify the
following user intent, accompanied by a simple response
draft.
Requirements:
{here are the given annotation requirements.}
Annotation formats:
{here are the explained annotation formats.}
Given annotated examples:
{here are some examples of already-annotated data.}
```

Listing 4: Instructions for annotators

## D Additional Demonstration

We further deliver a workflow knowledge showcase for 'flight booking' within the role 'flight\_inquiry' in the travel&transportation domain, accompanied by a generated cross-scenario ground-truth session.

**Text Format.** Below is the workflow knowledge embedded in text format for the scenario of flight booking under the role of flight\_inquiry.

```
The process of booking a flight begins when the user
initiates the request. The user is then prompted to provide
the flight ID. Once the flight ID is given, the system calls
the checkAvailability function to confirm the flight status
and check for availability based on the returned is_air
status.

If the flight is available, the system informs the user
about the flight's availability and requests their ID number
and name. The system then calls the reserveFlight function
to make the reservation and checks the success of the
reservation based on the returned is_successful status.

If the reservation is successful, the system informs the
user that the reservation succeeded and provides the booking
details, which are based on other returned values from the
reserveFlight. The user is then asked if they want to book
another flight.

If the user decides to book another flight, the process
loops back to requesting the user's ID number and name for
the new reservation. If the user does not want to book
another flight, they are welcomed to contact the system
again for future needs.

If the flight is not available, the system informs the user
that the flight is unavailable and asks if the user wants to
book another flight. If the user decides to book another
flight, the process returns to checking the availability of
the new flight. If the user does not wish to book another
flight, they are also welcomed to contact the system for
future needs.

If the reservation fails, the system notifies the user of
the failure and asks if they want to try booking another
flight. If the user agrees, the process loops back to
attempting the reservation again. If the user chooses not to
rebook, they are informed that they are welcome to contact
the system again in the future.
```

Listing 5: Text-Format workflow for flight booking

**Code Format.** Below is the workflow knowledge embedded in code format with python style for the scenario of flight booking.



```

def book_flight():
    request_initiated = True

    while request_initiated:
        # Prompt for flight ID
        flight_id = request_flight_id()

        # Check availability
        if check_availability(flight_id):
            # Inform availability and request user info
            inform_user_available()
            user_id = request_user_id()
            user_name = request_user_name()

            # Attempt reservation
            reservation_result = reserve_flight(flight_id,
            user_id, user_name)
            if reservation_result['is_successful']:
                # Inform success and provide details
                inform_reservation_success(
                reservation_result)

            # Ask if user wants another booking
            if user_wants_to_book_another():
                continue
            else:
                request_initiated = False
                inform_user_contact_again()
        else:
            # Inform failure and ask if user wants to
            try again
            inform_reservation_failure()
            if user_wants_to_try_again():
                continue
            else:
                request_initiated = False
                inform_user_contact_again()
        else:
            # Inform unavailability and ask if user wants to
            book another flight
            inform_user_unavailable()
            if user_wants_to_book_another():
                continue
            else:
                request_initiated = False
                inform_user_contact_again()

def request_flight_id():
    return input("Please provide the flight ID: ")

def check_availability(flight_id):
    # Simulate check availability (always returns True for
    this example)
    return True

def inform_user_available():
    print("The flight is available.")

def request_user_id():
    return input("Please provide your ID number: ")

def request_user_name():
    return input("Please provide your full name: ")

def reserve_flight(flight_id, user_id, user_name):
    # Simulate reservation (always succeeds for this example
    )
    return {'is_successful': True, 'flight_details': 'Flight
    _AA123_at_7:00AM_on_April_5'}

def inform_reservation_success(reservation_result):
    print("Reservation succeeded.")
    print(f"Booking details: {reservation_result}")

def user_wants_to_book_another():
    response = input("Do you want to book another flight?_(
    yes/no): ").strip().lower()
    return response == "yes"

def inform_user_contact_again():
    print("Thank you! Please contact us again for future
    needs.")

def inform_reservation_failure():
    print("Reservation failed.")

def user_wants_to_try_again():
    response = input("Do you want to try again?(yes/no): ")
    .strip().lower()
    return response == "yes"

def inform_user_unavailable():
    print("The flight is unavailable.")

# Start booking process
book_flight()

```

Listing 6: Code-Format workflow for flight booking

**Flowchart Format.** Below is the workflow knowledge embedded in flowchart format for the scenario of flight booking. Its visualization is also shown in Figure 7.

```

flowchart TD
    SK000(Start)--Book a flight-->SK001>Inquire the user for the
    Flight ID,
    call checkAvailability to confirm flight status,
    and check flight availability based on the returned is_air]
    SK001--Flight is available--> SK002>Tell the user that the
    flight is available.
    Ask the user for ID number and name, call reserveFlight, and
    check if the reservation is successful based on the
    returned is_successful ]
    SK002--Reservation succeeded-->SK003[Inform the user that
    the reservation succeeded, and notify the user of booking
    details based on other returned values of reserveFlight.
    Ask if the user wants to book again]
    SK002--Reservation failed-->SK004[Notify the user that the
    reservation failed and ask if the user wants to book again]
    SK004 & SK003--User books again-->SK002
    SK004 & SK003--User does not rebook-->SK006[The user is
    welcome to contact again for future needs]
    SK001--Flight is unavailable-->SK005[Inform the user that
    the flight is unavailable, and confirm if the user wants to
    book another flight]
    SK005--User books another flight-->SK001

```

Listing 7: Flowchart-Format workflow.

**Toolbox Information.** Below we provide the tool information within the knowledge base of the flight booking scenario. There are two available tools of *checkAvailability* and *reserveFlight*.

```

API: checkAvailability
API Description: Check the ticket availability for a flight
given by the user
Input parameters: {"plan_code": {"type": "string", "
description": "Flight ID, such as \CA1234, CZ5678, MU9101,
3U4567, FM8901, HU7890\", "required": true}, "
estimated_time": {"type": "string", "description": "
Estimated time, in the format: year + month + day + hour.
For example, \12 o'clock on January 7, 2199\". If not
filled in, the default is now.", "required": false}}
Output parameters: {"is_air": {"type": "string", "
description": "Whether the airplane can be ticketed.
Enumerated type, [true, false]}}

API: reserveFlight
API Description: Reserve the user's flight number and return
the reservation information
Input parameters: {"plan_code": {"type": "string", "
description": "Flight ID, such as \CA1234, CZ5678, MU9101,
3U4567, FM8901, HU7890\", "required": true}, "cabin_type":
{"type": "string", "description": "Cabin type. Enumerated
type, [Economy Class, Business Class, First Class]. If not
filled in, these types will be selected from front to back",
"required": false}, "estimated_time": {"type": "string", "
description": "Estimated time, in the format: year + month +
day + hour. For example, \12 o'clock on January 7, 2199\".
If not filled in, the default is now", "required": false}}
Output parameters: {"is_successful": {"type": "string", "
description": "Whether the reservation is successful.
Enumerated type, [true, false]}, "boarding_number": {"type
": "string", "description": "Boarding number, such as \C1,
D3, E7, F2, G4, H6\".}, "seat_number": {"type": "string", "
description": "Seat number, such as \12A, 15C, 21F, 28D, 34
B, 42E\".}, "airport_name": {"type": "string", "description
": "Airport name, such as \Beijing Capital International
Airport, Shanghai Pudong International Airport\".}, "
actual_time": {"type": "string", "description": "Actual time
, in the format: year + month + day + hour. For example,
\12 o'clock on January 7, 2199\". If not filled in, the
default is now.}}

```

Listing 8: Toolbox information example

**Session Example** Below we further provide a demonstration example of a cross-scenario ground-truth session for the role of flight\_inquiry, which includes the scenarios of flight booking and flight information search. As shown below, different interaction sources (user, agent, environment) are

indicated upfront. The agent may choose to reply with a 'Response' (interact with the user) or utilize a tool with 'Action' along with 'Action Input' (interact with the environment).

```
User:
Hello! I need to book a flight from New York to London for a
business meeting. Can you provide options for flight times,
duration, and pricing? Thank you!

Agent:
Response: Could you please provide me with the departure
date and time for your flight from New York to London?
Additionally, if you have any preferred airlines, please let
me know.

User:
Hi again! I'm looking to depart on April 5th, anytime in the
morning. No specific airline preferences; I'm just looking
for the best combination of price and travel time. Thanks!

Agent:
Response: Could you specify an approximate time in the
morning on April 5th for your departure? For example, would
around 7 AM work for you?

User:
Absolutely, around 7 AM would be perfect. Could you also
check if there are any flights available a bit later in the
morning, just in case? Thanks!

Agent:
Action: planeSearch
Action Input: {"starting_point": "New York", "end_point": "
London", "estimated_time": "7 o'clock on April 5, 2039"}

Environment:
{"plan_dict": [{"plan_code": "AA123", "plan_type": " Boeing
737", "departure_time": "April 5th, 2039, 7:00 AM", "
plan_price": "$500"}, {"plan_code": "BA456", "plan_type": "
Boeing 777", "departure_time": "April 5th, 2039, 8:00 AM", "
plan_price": "$550"}, {"plan_code": "DL789", "plan_type": "
Boeing 787 ", "departure_time": "April 5th, 2039, 9:00 AM", "
plan_price": "$530"}]}

Agent:
Response: I found some flights for you from New York to
London on the morning of April 5th. Here are the options:
1. Flight AA123: Departs at 7:00 AM, arrives at 7:00 PM,
Duration: 12 hours, Price: $500
2. Flight BA456: Departs at 8:00 AM, arrives at 8:00 PM,
Duration: 12 hours, Price: $550
3. Flight DL789: Departs at 9:00 AM, arrives at 9:00 PM,
Duration: 12 hours, Price: $530
Please let me know which flight you would prefer or if you
need more information on any of these options.

User:
Oh, thanks for the options! Could you tell me if any of
these flights offer Wi-Fi service or extra legroom? I need
to work during the flight.

Agent:
Response: I apologize, but I don't have information
regarding specific amenities like Wi-Fi service or extra
legroom on flights. You might want to check directly with
the airlines for detailed service offerings. Can I assist
you with anything else related to your flight booking?

User:
That's understandable, thanks! Could you please help me book
Flight AA123 then? I'll contact the airline directly for
the extra details.

Agent:
Action: checkAvailability
Action Input: {"plan_code": "AA123", "estimated_time": "7 o'
clock on April 5, 2039"}

Environment:
{"is_air": "true"}

Agent:
Response: The flight AA123 on April 5th at 7:00 AM is
available. To book this flight, could you please provide
your ID number and full name as it appears on your
identification?

User:
Sure! My full name is Johnathan L. Smith and my ID number is
987654321. Please proceed with the booking. Thanks for your
help!

Agent:
Action: reserveFlight
Action Input: {"plan_code": "AA123", "estimated_time": "7 o'
clock on April 5, 2039"}
```

```
Environment:
{"is_successful": "true", "boarding_number": "C1", "
seat_number": "12A", "airport_name": "John F. Kennedy
International Airport", "actual_time": "7 o'clock on April
5, 2039"}

Agent:
Response: Your reservation for Flight AA123 on April 5th at
7:00 AM from John F. Kennedy International Airport has been
successfully booked. Here are your booking details:
- Boarding number: C1
- Seat number: 12A
- Airport: John F. Kennedy International Airport
- Departure Time: 7 o'clock on April 5, 2039
Would you like to book another flight or is there anything
else I can assist you with?

User:
No, that's all for now. Thanks for all your help!

Agent:
Response: You're welcome, Johnathan! If you need any more
assistance in the future, feel free to reach out. Have a
great trip and a successful business meeting in London! Safe
travels!
```

Listing 9: Ground-truth cross-scenario session example