PlanningArena: A Modular Benchmark for Multidimensional Evaluation of Planning and Tool Learning

Zihan Zheng[1*], **Tianle Cui**[1*], **Chuwen Xie**[1],
Jiahui Pan[1], **Qianglong Chen**[2†], **Lewei He**[1†]
[1]South China Normal University [2]Zhejiang University
chenqianglong@zju.edu.cn helewei@m.scnu.edu.cn

Abstract

One of the research focuses of large language models (LLMs) is the ability to generate action plans. Recent studies have revealed that the performance of LLMs can be significantly improved by integrating external tools. Based on this, we propose a benchmark framework called PlanningArena, which aims to simulate real application scenarios and provide a series of apps and API tools that may be involved in the actual planning process. This framework adopts a modular task structure and combines user portrait analysis to evaluate the ability of LLMs in correctly selecting tools, logical reasoning in complex scenarios, and parsing user information. In addition, we deeply diagnose the task execution effect of LLMs from both macro and micro levels. The experimental results show that even the most outstanding GPT-4o and DeepSeekV3 models only achieved a total score of 56.5% and 41.9% in PlanningArena, respectively, indicating that current LLMs still face challenges in logical reasoning, context memory, and tool calling when dealing with different structures, scenarios, and their complexity. Through this benchmark, we further explore the path to optimize LLMs to perform planning tasks.

Figure 1: Illustrations of a simple travel planning task.

1 Introduction

In recent years, large language models (LLMs) (Hurst et al., 2024; Yang et al., 2024; Team et al., 2024a; Shao et al., 2024b; Dubey et al., 2024; Jiang et al., 2023) have demonstrated remarkable capabilities in tool utilization (Qin et al., 2024; Shen et al., 2024c; Ma et al., 2024a; Liu et al., 2024) and task planning (Wang et al., 2023a; Hao et al., 2023; Glória-Silva et al., 2024). As shown in the figure 1, integrating external tools through feedback mechanism (Wu et al., 2024b) and retrieval-augmented generation (RAG) techniques (Huang et al., 2024b; Kong et al., 2024; Lee et al., 2024) with external tools (Yao et al., 2023b; Shi et al., 2024) has emerged as a pivotal approach to enhancing model performance. Despite advancements in planning architectures and learning methodologies (Guo et al., 2024b; Sun et al., 2024), the evaluation of LLMs' tool planning capabilities (Qin et al., 2023; Zheng et al., 2024b; Shen et al., 2024b; Zhang et al., 2024a,b) presents fundamental challenges.

Existing evaluation systems face four critical limitations: 1) Domain-specific benchmarks struggle to capture cross-scenario planning abilities (Shao et al., 2024a; Zheng et al., 2024a; Xie et al., 2024); 2) API-centric evaluation frameworks often fail to adequately reflect real-world application ecosystems (Basu et al., 2025; Li et al., 2023; Guo et al., 2024c; Basu et al., 2024; Huang et al., 2024a); 3) Current datasets do not simulate complex real-world tasks and overlook task dependencies (Yin

[*]Equal contribution.
[†]Corresponding Authors.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31047–31086

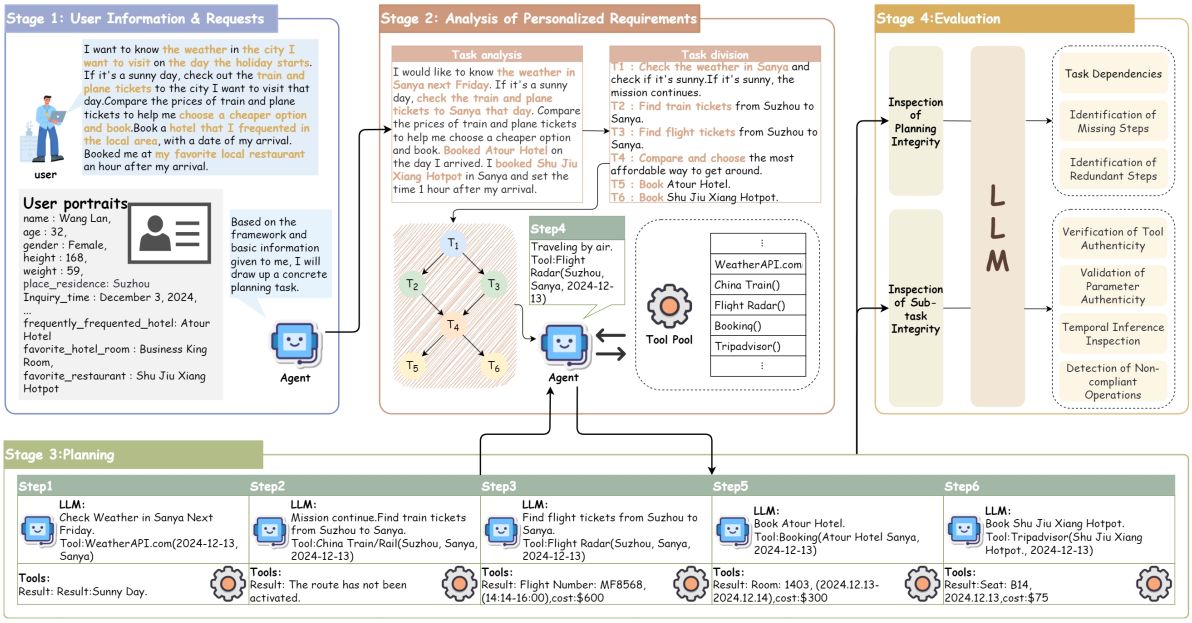

Figure 2: The pipeline of PlanningArena Benchmark. In-breadth, we analyze three stages (user information and requests, analysis of personalized requirements, planning) that could influence the planning process from the perspective of user needs. We employ an in-depth, multi-level evaluation (task analysis, tool selection, task execution, and evaluation) to diagnose the reasons for potential issues in LLM-based planning.

et al., 2024; Wang et al., 2024a; Shen et al., 2024d; Zhang et al., 2024a); 4) Static scenario constructions neglect dynamic user needs (Sun et al., 2025; Tan and Jiang, 2023).

This paper introduces PlanningArena, a comprehensive dataset designed specifically for planning and tool-based task design, addressing the aforementioned issues through three innovations: First, it integrates 10 real-world scenarios with API/APP composite workflows; second, it establishes five types of planning structures—Single-APP, Cross-APP, parallel independent, chain-dependent, and directed acyclic graphs; and third, it employs a multi-agent data synthesis framework (Guo et al., 2024a; Hong et al., 2024; Chen et al., 2024; Wang et al., 2023b; Chang et al., 2024) and a dynamic update mechanism to prevent data contamination (Jiang et al., 2024). A representative example is illustrated in Figure 2.

Experimental analysis of 10 LLMs (5 commercial and 5 open-source models) reveals two key findings: 1) All models exhibit significant performance degradation as task complexity increases; 2) Performance variability is pronounced in multi-context planning scenarios. These results highlight the inherent limitations of current LLM architectures in complex tool orchestration.

The primary contributions of this study include:

1) We construct a novel evaluation platform combining real-world scenario simulation with structured task decomposition.

2) We provide a comprehensive empirical analysis of LLM planning capabilities across multiple complexity dimensions.

3) We propose an extensible data synthesis paradigm integrating multi-agent generation and user profile modeling.

The dataset generation code and evaluation protocols are open-sourced at: `https://github.com/KeLes-Coding/PlanningArena`.

2 Related Work

2.1 Tool Learning

As the performance of large language models (LLMs) rapidly improves, it is foreseeable that LLMs will gradually develop capabilities akin to humans in using tools to solve complex problems (Qin et al., 2024; Parisi et al., 2022; Ji et al., 2023), a concept referred to as tool learning (Qu et al., 2024). By incorporating external tools for interaction with LLMs, not only can their inherent limitations be effectively mitigated (Tang et al., 2023), but it also facilitates dynamic knowledge acquisition and integration (Nakano et al., 2022; Komeili et al., 2022; Zhang et al., 2023), thereby providing more precise and contextually relevant outputs. Furthermore, the application of domain-specific tools can enhance LLMs' specialized knowledge, improving their practicality in professional settings (Kadlčík et al., 2023; He-Yueya et al., 2023; Chen et al., 2021).

2.2 Planning Benchmark

Planning ability, a hallmark of human intelligence, involves complex processes (Ghallab et al., 2004). As LLMs advance, research on their planning capabilities has deepened (Zuo et al., 2025; Sirdeshmukh et al., 2025; Liu et al., 2023; Valmeekam et al., 2023; Wei et al., 2023), covering task decomposition (Shen et al., 2023), plan selection (Wang et al., 2023a; Yao et al., 2023a; Gao et al., 2024; Hao et al., 2023), and external module assistance (Liu et al., 2023; Dagan et al., 2023; Guan et al., 2023). RAG enhances long-term conversational coherence through context storage and retrieval (Lewis et al., 2020; Mao et al., 2021).

Previous research has significantly advanced the development of evaluation tools and methodologies in multiple domains (Ma et al., 2024b; Farn and Shin, 2023; Ruan et al., 2024), particularly in the evaluation of planning and reasoning capabilities (Shen et al., 2024a; Wang et al., 2024b). The tool planning process is subdivided into four critical stages: recognizing necessity (Huang et al., 2024c; Ning et al., 2024), task planning (Xu et al., 2023; Wu et al., 2024a), tool selection (Song et al., 2023; Huang et al., 2024b), and tool invocation (Qin et al., 2023; Ye et al., 2024). Current benchmarking efforts predominantly focus on evaluating large language models' (LLMs) abilities at these different stages.

Despite contributions from existing tools to LLM evaluation, limitations remain: narrow scenario simulation (Xie et al., 2024; Zheng et al., 2024a), questionable real-world applicability of some research tools (Patil et al., 2023; Schick et al., 2023), and neglected task dependencies (Shen et al., 2024d; Styles et al., 2024; Zhang et al., 2024a). To address these, we introduce PlanningArena—a platform for realistic scenario simulation that emphasizes task dependencies, aiming for a more authentic and comprehensive assessment of planning capabilities.

3 Design

In this section, we introduce the construction of the PlanningArena benchmark, including the design principles and the scenarios concerned. Unlike previous benchmarks, we emphasize daily scenarios to evaluate the planning ability of LLMs in terms of tool using based on common knowledge.

3.1 Design Principles

To comprehensively assess the planning abilities of LLMs, we evaluate them with the following principles:

- Breadth and Reality: We focus on a wide range of real-world scenarios to assess general common knowledge-based planning.
- Depth and Dynamics: We emphasize multi-round interactions to evaluate dynamic planning capabilities.

3.1.1 Breadth and Reality

In real-world scenarios, user query is often colloquial, and the planning task is more complex and relies on the model's implicit commonsense reasoning capabilities based on user preferences. While existing benchmarks fail to reflect real-world challenges, we introduce PlanningArena, which covers diverse real-life planning scenarios to access implicit reasoning, including temporal reasoning (such as date calculation, special days recognization and schedules adjustment) and personalized reasoning. More details can be found in Appendix A.1. Given that different scenarios require different planning abilities and different common knowledge, we consider ten scenarios in PlanningArena to evaluate the planning ability comprehensively; more details can be found in Appendix A.2.

Considering the impact of personalized information, we build a user profile database to evaluate the performance of models in handling personalized requirements and using user preferences.

3.1.2 Depth: multi-round planning LLM performance evaluation

Complex task planning requires multiple rounds of iterative guidance, especially when task parameters are interdependent, and we designed a two-tiered testing framework that progressively drills down from macro to micro to ensure a comprehensive assessment. The framework consists of the following two levels:

(1) **Planning Integrity Detection.**

In the first phase of the evaluation, we check whether the planning generated by LLM meets the key metrics at the macro level and subdivide the tasks to verify the completion of subtasks and dependencies.

Step Correctness. In LLM planning, $PlanTask$ is divided into subtasks (st) $\{st_1, st_2, ..., st_n\}$ with specific goals. The Step Correctness Assessment focuses on verifying that each step achieves the desired result.

Step Execution Rate. From a macro-perspective, this phase evaluates LLM planning coverage of subtasks defined in the original $PlanTask$. To measure Step Execution Rate (SER), we introduce a quantitative metric, calculated as:

$$SER = \frac{\sum_{i=1}^{n_{st}} w_i \cdot c_i \cdot q_i}{\sum_{i=1}^{n_{st}} w_i} \tag{1}$$

where n_{st} is the number of total subtasks; w_i is the ith subtask's importance; c_i indicates completion ($c_i = 1$ for completed, $c_i = 0$ otherwise); and q_i is the quality of completion in the range $[0, 1]$.

Accuracy verification of dependencies. In PlanningArena, the core test focus is on LLM's handling of task dependencies. We design a metric Dependency Accuracy (DA):

$$DA = \frac{m}{n} \sum_{i=1}^{m} \left(\prod_{j=1}^{|B_i|} (1 - B_{ij}) \cdot w_{ij} \right) \tag{2}$$

n is the total dependencies; m is LLM-generated dependencies; B_i is the Boolean set for the ith dependency; $B_{ij} = 1$ indicates "hallucination"; $|B_i|$ is checkpoint count; w_{ij} is checkpoint weight; N_{B_i} is total checkpoints.

Combined evaluation of these aspects determines task success or step redundancy from incorrect dependencies. By multiplying SER with DA, we compute the **Logical Pass Rate (LPR)** of the planning task as a quantitative assessment of whether the plan execution follows the intended task logic.

(2) **Subtask Integrity Detection.**

In PlanningArena, the second phase evaluates the occurrence of fictitious tools and parameter matching by subdividing tasks, verifying the correct use of API and APP tools, and the illusion of parameters.

We define Sub-mission Adoption Rate (SMAR) integrity detection metrics:

$$SMAR = \begin{cases} \dfrac{\sum_{i=1}^{|B|} w_i \cdot (1 - B_i)}{N_B}, & \text{if } B_{gc} = 0 \\ 0, & \text{if } B_{gc} = 1 \end{cases}$$

where B_{gc} indicates garbled characters; B is a set of Boolean indicators $\{B_{MS}, B_{RS}, B_{FP}, B_{FT}, B_{TIE}\}$, corresponding to the checkpoints for missing steps, redundant steps, fictitious parameters, fictitious Tools, and time inference errors.

4 Data Construction

Given user query and user profile, we mainly focus on APP based Planning and API based Planning, where the task complexity increases from single tools to multi-tools dependencies.

4.1 APP-based Planning

In App based planning, we mainly evaluate how LLMs understand the function of App, such as DiDi Chuxing, Amazon, or others.

Single-APP (SAPP) Planning. The single-APP planning evaluates LLMs' capacity to achieve user goals through atomic mobile operations (e.g., tap, text input, swipe). This framework assesses two core competencies: (1) task decomposition using native app functionalities, and (2) execution proficiency within constrained operational primitives (back/home navigation, gesture recognition). These metrics reveal LLMs' potential for real-world mobile interaction capabilities.

Cross-APP (CAPP) Planning. Cross-APP planning evaluates the parallel and sequential task orchestration capabilities of LLMs across multiple interfaces. This scenario is based on single-application testing and further verifies three key dimensions: (1) cross-platform consistency of workflow execution, (2) dynamic parameter adaptation in complex environments, and (3) context-aware personalization with user profile fusion. Successful implementation requires simultaneous processing of semantic understanding of operation sequences and personalized behavior patterns.

4.2 API-based Planning

We divide the API based planning into three levels, namely parallel, chained and DAG types, with progressively increasing complexity of the task:

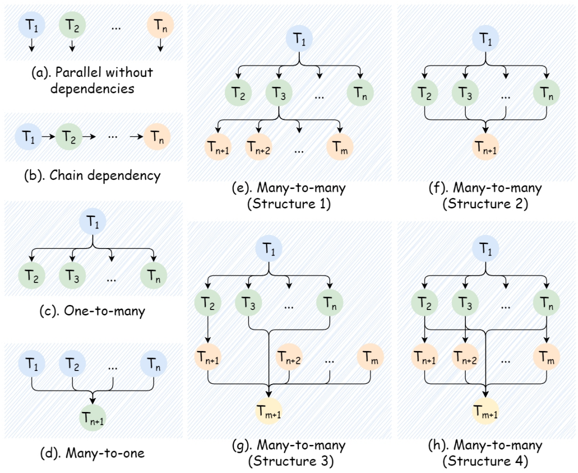

Figure 3: The diagram illustrates various dependency structures in task execution workflows. Specifically, it depicts (a) Parallel tasks without dependencies, (b) Chain of dependent tasks, (c) One-to-many dependencies, (d) Many-to-one dependencies, and (e-h) different configurations of Many-to-many dependencies.

Parallel API (PAPI) Planning. As shown in Figure 3.a, we design parallel tasks without any dependencies, to evaluate the performance of LLM when processing multiple independent tool calling simultaneously.

Chained API (CAPI) Planning. As shown in Figure 3.b, we evaluate the ability of LLMs to manipulate the pre- and post-placement API when dealing with linear chained dependencies.

DAG Based API (DAPI) Planning. Since the structure of DAG can vary significantly, we divide it into three sub-structures, including one-to-many (3.c), many-to-one (3.d), and many-to-many (3.e-h). We provide a detailed description of the substructures of DAG in A.3.

Ultimately, each data sample includes the following components as shown in A.4.

4.3 User Profile Personalization

In the process of building PlanningArena, we introduce user profile as an important consideration, which is different from traditional benchmarking methods. In order to systematically build user portraits, we design a set of user templates, which set 11 detailed information dimensions for the tool chain corresponding to each planning scenario. These dimensions not only include the basic physiological characteristics of users (such as height, weight, age, etc.), but also extend to preference settings related to ten specific planning scenarios, such as hotel booking habits, budget, and travel mode selection in the travel scenario; clothing purchase preferences, brand loyalty, and channel selection in the shopping scenario; past medical conditions, exercise habits, and nutritional needs in the health maintenance scenario; subject interests and learning aids in the education scenario; and entertainment forms and consumption venues in the entertainment scenario. On the one hand, these preference settings match the input parameters of the tool chain, avoiding parameter missing caused by human factors; on the other hand, they enrich the information volume of planning tasks, making these tasks closer to real scenarios.

Based on the above template framework, we use LLM to generate 50 user portraits with unique attributes. Each portrait reflects the personalized characteristics and demand patterns of different users in multiple planning scenarios. Specific user portrait examples are detailed in the Appendix A.5.

We then provide LLMs with personal information and tasks based on these profiles, asking the model to analyze user needs and task dependencies independently and develop a planning scheme (detailed in Appendix A.6). We demonstrate the influence of adding User Profiles to the LLM planning performance in the AppendixA.7.

4.4 Data Synthesis Pipeline

Since manual data collection is costly, we use an automated data construction pipeline, combined with manual review and iterative correction, to ensure the quality, diversity and availability of data. Data construction mainly includes seed data construction and iterative data synthesis.

Scenario Selection. The scenario selection process is guided by two main criteria:

- The frequency of occurrence of planning task scenarios in real-world environments.

- The logical coherence and contextual relevance within the task chain. Based on these criteria, we collected over 100 subtasks from various domains, including e-commerce, transportation, and educational resources.

By combining and filtering these subtasks, we extracted the ten most comprehensive planning scenarios: Travel, Shopping, Entertainment, Development, Diet, Health, Education, Meeting, Game, and Calendar.

Tool Chain Generation. We collect more than 16,000 app interfaces and 150 APP tools through Rapid[1] and Google Play[2]. Then, we build more than 200 task execution chains (ToolChains) to provide a robust infrastructure for task planning and tool integration, ensuring that all tasks can be implemented with the tools provided by PlanningArena. Taking the travel scenario as an example, users can query the weather at the destination through Weather API, book air tickets using FlightRadar API, and finally complete hotel reservations with the help of Booking API, thus forming a coherent task execution process.

Seed Task Construction. Based on the pre-designed Tool Chains, we manually construct 300 seed planning tasks and strictly screen the data in the process. While incorporating the task structure described in Section 4, we ensure the accuracy of the logical dependencies within each task. It is worth noting that although these seed tasks rely on specific tool chains, most tasks have multiple feasible solutions thanks to the diversity of tools.

Command Evolution. Inspired by (Mitra et al., 2024; Wang et al., 2023b) et al. We employ a multi-agent based approach to automate the production of derived data in the data evolution phase. This approach utilizes seed tasks as the original data source, and iteratively derives datasets with high diversity and varying complexity through Agents. We present the implementation details of Command Evolution in Appendix A.9.

4.5 Data Statistics

Table 1: Statistics of PlanningArena.

	User	Difficulty level			Overall
		Easy	Middle	Hard	
Total	50	1703	1486	1311	4500
Single-APP	20	730	371	99	1200
Cross-APP		313	333	554	1200
PAPI		280	162	158	600
CAPI		80	320	200	600
DAPI	30	300	300	300	900

Table 1 presents the statistical analysis of the PlanningArena benchmark, which contains 4,500 data samples and 50 user profiles. The samples are divided into five groups based on the task structure complexity and tool usage type: SAPP, CAPP, PAPI, CAPI, and DAPI. Each group is further divided into three difficulty levels: easy, medium, and hard according to task length, structure complexity, number of tools, and scenario difficulty score. This classification method provides a systematic framework for evaluating the planning performance of different models in various task environments. Specific details for each difficulty level can be found in the Appendix A.10.

5 Experiments

5.1 Baselines

We utilize 10 current mainstream LLMs, including five proprietary models and five open-source models (their details can be viewed in Appendix B.1). Among them, the five proprietary models include Gemini-1.5 series (Gemini-1.5-flash and Gemini-1.5-pro) (Team et al., 2024a), GPT-4o series (GPT-4o and GPT-4o-mini) (Hurst et al., 2024) and Qwen-plus (Qwen et al., 2025). Five currently advanced open source models include: Llama3.1 series (Llama3.1-8B and Llama3.1-70B) (Dubey et al., 2024), Deepseek-V3 (DeepSeek-AI et al., 2024), Gemma-2-9B (Team et al., 2024b) and GLM-4-9B (GLM et al., 2024).

5.2 Evaluation Settings

In order to ensure the consistency and reproducibility of the results, we develop specialized calling configurations for different models. For proprietary models and DeepSeek-V3, official APIs are used for calling; for other open source models, the inference environment is built based on the Ollama framework. All models are deployed on three NVIDIA RTX 4090 GPUs. In the experiments, the temperature parameter remains 0.0 and the rest of the parameters are kept at their default values to minimize the variable effects.

5.3 Main Results

As shown in the table 2, we demonstrate the overall performance of different LLMs on PlanningArena. The tested metrics include Single-APP (SAPP), Cross-APP (CAPP), Parallel API (PAPI), Chained API (CAPI), Dag API (DAPI) with one-to-many (OM), many-to-one (MO), and many-to-many (MM) DAG structures.

In the comprehensive performance evaluation, GPT-4o achieves a correctness rate of 56.5%, lead-

[1]https://rapidapi.com
[2]https://play.google.com

Table 2: Performance of different LLMs on PlanningArena.

Model	APP			API										Overall
	SAPP	CAPP	Overall	PAPI	CAPI	DAPI	OM	MO	MM1	MM2	MM3	MM4	Overall	
Proprietary														
GPT-4o	**65.6**	**67.8**	**66.7**	**44.4**	40.0	**44.4**	**54.6**	**61.2**	**58.4**	**45.4**	28.7	**18.6**	**44.4**	**56.5**
GPT-4o-mini	11.7	17.2	14.4	26.7	30.0	18.9	15.3	31.2	0.0	26.8	26.4	13.6	25.2	19.1
Gemini-1.5-flash	53.9	33.9	43.9	43.3	**41.1**	12.2	0.0	6.7	13.6	40.0	12.3	0.0	31.5	38.9
Gemini-1.5-pro	55.6	35.6	45.6	37.8	32.2	13.3	20.0	26.7	20.0	8.7	4.7	0.0	27.8	37.9
Qwen-plus	46.3	47.0	46.7	38.5	32.8	28.6	44.7	25.3	30.7	37.3	29.3	4.0	32.6	40.1
Open-weight														
DeepSeekV3	45.5	47.4	46.5	37.2	34.8	37.6	43.3	49.3	39.3	36.7	**39.3**	17.3	36.7	41.9
Gemma-2-9B	32.2	23.3	27.8	12.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.1	17.6
GLM-4-9B	6.1	4.4	5.3	6.7	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	4.3
Llama-3.1-8B	13.3	17.8	15.6	17.8	13.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.4	13.3
Llama-3.1-70B	32.2	36.1	34.2	23.3	18.9	27.8	6.7	33.3	**60.0**	33.3	26.7	6.7	23.3	29.5

ing all competitors in terms of tool-enhanced planning capabilities, and realizes a 34.8% performance improvement compared to DeepSeek-V3, which comes in second. DeepSeek-V3 performs particularly well in the open source modeling domain and outperforms all proprietary models except GPT-4o.

For the APP tool planning task, GPT-4o demonstrates high performance in both SAPP and CAPP scenarios, and we note that the balanced performance of the top three APP planning performing models (GPT-4o, DeepSeekV3, and Qwen-plus) in both of these scenarios prove their robustness in single-task execution and cross-task coordination. Although the Gemini series performs well in the SAPP task, its CAPP task performance drops by more than 15%, showing a lack of cross-scenario generalization capability.

In the API tool planning task, GPT-4o demonstrates its capability to handle multi-task branching (PAPI) and long-range dependency modeling (CAPI). Notably, Gemini-1.5-flash exhibits better performance than the best in both, except for GPT-4o. While GPT-4o and DeepSeekV3 score higher on MM1 and MM2, the significant drops on MM3 and MM4 reveal the model's limitations in dealing with deep nested relationships. The plummeting performance of the Gemini-1.5 series in the MM task further emphasizes its deficiencies in complex nested relationship construction and cross-scenario generalization capabilities, while other models (e.g., the LLama-3.1-8b) generally scored close to zero in the MM task, highlighting their deficiencies in modeling complex dependencies.

In addition, we show the detailed results of the three best performing models: GPT-4o, DeepSeekV3, and Qwen-plus in Appendix B.2. For detailed planning procedures, refer to Appendix B.3.

5.4 Error Analysis

5.4.1 Error analysis in difficulty level

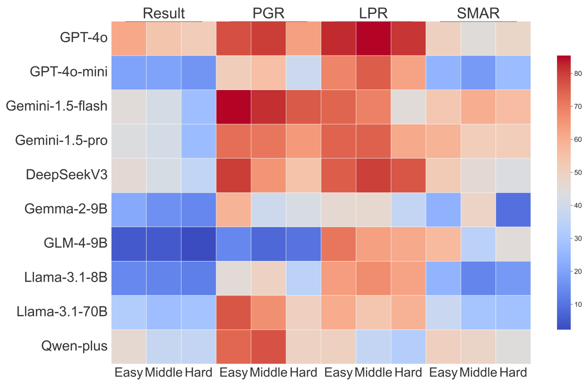

Figure 4: Comparative analysis of different models under various evaluation indicators, including results, plan generation rate (PGR), Logic pass rate (LPR), and Submission adoption rate (SMAR).

As shown in Figure 4, this study conducts a comparative analysis of the performance of multiple models in planning tasks of different complexity and comes to the following conclusions:

Task text complexity affects plan generation. PGR is negatively correlated with the task text length, and this phenomenon is especially obvious in the "difficult" level (average number of tokens in the task 191.8) tasks. The significant decrease in PGR for most models in the difficult task group indicates that current language models still have limitations in handling long context tasks.

Task structural complexity constrains PLAN parsing accuracy. In PlanningArena, the number of subtasks is positively correlated with the task structure complexity, which makes the model encounter more challenges in parsing complex tasks, often resulting in missing subtasks or incorrectly judged subtask dependencies. This reveals the limitations of existing models in understanding complex task structures.

Information Density Interference Tool Selection Mechanism. As task complexity increases, LLM inevitably needs to invoke more tools to realize the planning task, which makes it necessary for the model to accurately extract information from higher information density sources (e.g., information about the task itself and user portraits). This leads to an increase in the frequency of tool invocation errors and the presence of user information mismatches during invocations, suggesting that existing models lack reliable contextual tool-parameter mapping mechanisms for complex planning tasks.

Based on the above analysis, we believe that future research should focus on building a graph-based task representation learning framework to improve the generalization ability of intelligent agents in open environments by modeling topological dependencies between complex tasks.

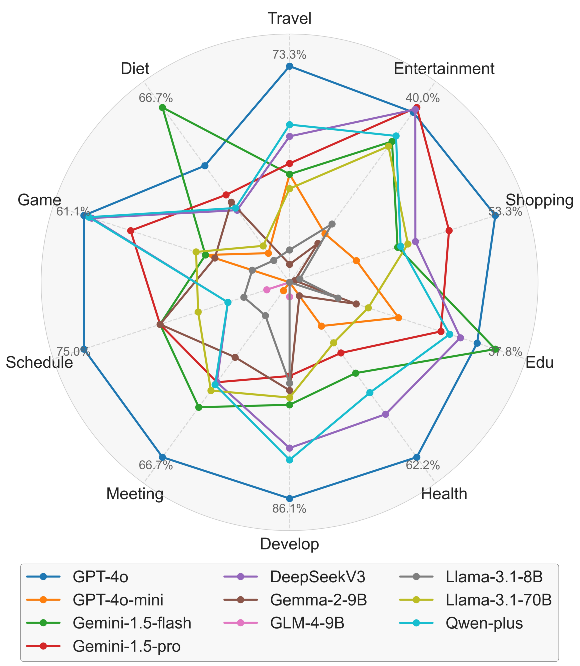

Figure 5: Comparison of the performance of different models in different scenarios. The performance of each model is represented by the line connecting the points on the chart, and the percentage value indicates the performance level in each field.

5.4.2 Error analysis in different scenarios

As shown in Figure 5, we evaluate the task completion of the test model in different application scenarios. The results show that in scenarios involving a large amount of user information processing (such as travel, entertainment, and shopping), the model needs to accurately parse specific task requirements based on user portraits, but in these data-intensive scenarios, the model shows a high parameter configuration error rate and insufficient environmental adaptability.

In addition, in scenarios related to time reasoning (such as travel, entertainment, and diet), the model's temporal logic processing capabilities are limited. Even the optimal GPT-4o model has a temporal reasoning accuracy of only 65.7%, which significantly affects the performance of LLM in time-related tasks.

For task scenarios with complex logic dependencies, such as ordering tasks in diet, the model is required to perform constraint queries for specific restaurants or dishes (such as checking ratings and user reviews) throughout the task cycle and implement cross-platform price comparisons. When faced with such complex structures and numerous subtasks, the step missing rate of LLM increases significantly, revealing its performance limitations in personalized demand parsing, temporal reasoning, and long-chain task logic processing.

The above analysis reveals the performance limitations of LLM in processing complex user information, temporal reasoning, and long-chain logical tasks, suggesting that future improvements should focus on enhancing the model's ability to parse personalized requirements, improving the accuracy of temporal reasoning, and optimizing the processing mechanism of complex logical tasks.

6 Conclusion

This paper introduces PlanningArena Benchmark, a dynamic dataset designed to evaluate the planning capabilities of LLMs when utilizing tools. By integrating user profiles and modular tasks, PlanningArena constructs personalized and structurally complex task sets to assess LLM performance across diverse planning scenarios. Our multi-stage evaluation framework, ranging from macro to micro levels, provides a comprehensive assessment of LLMs' planning and tool learning abilities. The results indicate that current LLMs still face significant challenges in logical parsing and contextual reasoning when using tools and planning tasks. PlanningArena not only advances personalized benchmarking for tool planning but also sets a new standard for the development of tool-augmented LLMs. Future work will focus on addressing the limitations identified and further enhancing the benchmark's capabilities.

7 Limitations

Despite its significant contributions, PlanningArena has several limitations that highlight areas for future improvement. First, the current version of PlanningArena is primarily limited to textual planning tasks. Future work will aim to extend the benchmark to multi-modal evaluation, covering a broader range of domains such as mobile application interaction, video content parsing, and audio information extraction. This expansion will enhance the benchmark's applicability and provide a more comprehensive evaluation of LLMs in real-world scenarios. Second, given the complexity of the dataset, manual evaluation is impractical, and we currently rely on automated evaluation using LLMs. While this approach ensures scalability, it may introduce systematic biases. To address this, future work will explore more robust evaluation mechanisms to improve assessment accuracy and reduce potential biases.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 52308250, the STI 2030-Major Projects under grant 2022ZD0208900, and the Guangdong Basic and Applied Basic Research Foundation under grant 2023A1515140100.

References

Kinjal Basu, Ibrahim Abdelaziz, Subhajit Chaudhury, Soham Dan, Maxwell Crouse, Asim Munawar, Vernon Austel, Sadhana Kumaravel, Vinod Muthusamy, Pavan Kapanipathi, and Luis Lastras. 2024. API-BLEND: A comprehensive corpora for training and benchmarking API LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12859–12870, Bangkok, Thailand. Association for Computational Linguistics.

Kinjal Basu, Ibrahim Abdelaziz, Kiran Kate, Mayank Agarwal, Maxwell Crouse, Yara Rizk, Kelsey Bradford, Asim Munawar, Sadhana Kumaravel, Saurabh Goyal, Xin Wang, Luis A. Lastras, and Pavan Kapanipathi. 2025. Nestful: A benchmark for evaluating llms on nested sequences of api calls. *Preprint*, arXiv:2409.03797.

Hsin-Yu Chang, Pei-Yu Chen, Tun-Hsiang Chou, Chang-Sheng Kao, Hsuan-Yun Yu, Yen-Ting Lin, and Yun-Nung Chen. 2024. A survey of data synthesis approaches. *Preprint*, arXiv:2407.03672.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2024. Autoagents: A framework for automatic agent generation. *Preprint*, arXiv:2309.17288.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Gautier Dagan, Frank Keller, and Alex Lascarides. 2023. Dynamic planning with a llm. *Preprint*, arXiv:2308.06391.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nicholas Farn and Richard Shin. 2023. Tooltalk: Evaluating tool-usage in a conversational setting. *Preprint*, arXiv:2311.10775.

Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. 2024. Interpretable contrastive monte carlo tree search reasoning. *Preprint*, arXiv:2410.01707.

Malik Ghallab, Dana Nau, and Paolo Traverso. 2004. *Automated planning. Theory practice*.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Diogo Glória-Silva, Rafael Ferreira, Diogo Tavares, David Semedo, and Joao Magalhaes. 2024. Plan-grounded large language models for dual goal conversational settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1271–1292, St. Julian's, Malta. Association for Computational Linguistics.

Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Preprint*, arXiv:2305.14909.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024a. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Yiduo Guo, Yaobo Liang, Chenfei Wu, Wenshan Wu, Dongyan Zhao, and Nan Duan. 2024b. Learning to plan by updating natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10062–10098, Miami, Florida, USA. Association for Computational Linguistics.

Zishan Guo, Yufei Huang, and Deyi Xiong. 2024c. CToolEval: A Chinese benchmark for LLM-powered agent evaluation in real-world API interactions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15711–15724, Bangkok, Thailand. Association for Computational Linguistics.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Joy He-Yueya, Gabriel Poesia, Rose E. Wang, and Noah D. Goodman. 2023. Solving math word problems by combining language models with symbolic solvers. *Preprint*, arXiv:2304.09102.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.

Shijue Huang, Wanjun Zhong, Jianqiao Lu, Qi Zhu, Jiahui Gao, Weiwen Liu, Yutai Hou, Xingshan Zeng, Yasheng Wang, Lifeng Shang, Xin Jiang, Ruifeng Xu, and Qun Liu. 2024a. Planning, creation, usage: Benchmarking LLMs for comprehensive tool utilization in real-world complex scenarios. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4363–4400, Bangkok, Thailand. Association for Computational Linguistics.

Tenghao Huang, Dongwon Jung, Vaibhav Kumar, Mohammad Kachuee, Xiang Li, Puyang Xu, and Muhao Chen. 2024b. Planning and editing what you retrieve for enhanced tool learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 975–988, Mexico City, Mexico. Association for Computational Linguistics.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2024c. Metatool benchmark for large language models: Deciding whether to use tools and which to use. *Preprint*, arXiv:2310.03128.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. Investigating data contamination for pre-training language models. *Preprint*, arXiv:2401.06059.

Marek Kadlčík, Michal Štefánik, Ondrej Sotolar, and Vlastimil Martinek. 2023. Calc-X and calcformers: Empowering arithmetical chain-of-thought through interaction with symbolic systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12101–12108, Singapore. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Yilun Kong, Jingqing Ruan, YiHong Chen, Bin Zhang, Tianpeng Bao, Shi Shiwei, du Guo Qing, Xiaoru Hu, Hangyu Mao, Ziyue Li, Xingyu Zeng, Rui Zhao, and Xueqian Wang. 2024. TPTU-v2: Boosting task planning and tool usage of large language model-based agents in real-world industry systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 371–385, Miami, Florida, US. Association for Computational Linguistics.

Myeonghwa Lee, Seonho An, and Min-Soo Kim. 2024. Planrag: A plan-then-retrieval augmented generation for generative large language models as decision makers. *Preprint*, arXiv:2406.12430.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. API-bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116, Singapore. Association for Computational Linguistics.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. Llm+p: Empowering large language models with optimal planning proficiency. *Preprint*, arXiv:2304.11477.

Hao Liu, Zi-Yi Dou, Yixin Wang, Nanyun Peng, and Yisong Yue. 2024. Uncertainty calibration for tool-using language agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16781–16805, Miami, Florida, USA. Association for Computational Linguistics.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, and Aixin Sun. 2024a. SciAgent: Tool-augmented language models for scientific reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15701–15736, Miami, Florida, USA. Association for Computational Linguistics.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, and Weizhu Chen. 2024b. Sciagent: Tool-augmented language models for scientific reasoning. *Preprint*, arXiv:2402.11451.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei ge Chen, Olga Vrousgos, Corby Rosset, Fillipe Silva, Hamed Khanpour, Yash Lara, and Ahmed Awadallah. 2024. Agentinstruct: Toward generative teaching with agentic flows. *Preprint*, arXiv:2407.03502.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *Preprint*, arXiv:2112.09332.

Kangyun Ning, Yisong Su, Xueqiang Lv, Yuanzhe Zhang, Jian Liu, Kang Liu, and Jinan Xu. 2024. Wtu-eval: A whether-or-not tool usage evaluation benchmark for large language models. *Preprint*, arXiv:2407.12823.

Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *ArXiv*, abs/2205.12255.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *Preprint*, arXiv:2305.15334.

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. 2024. Tool learning with foundation models. *Preprint*, arXiv:2304.08354.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, 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. *Preprint*, arXiv:2307.16789.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2024. Tool learning with large language models: A survey. *arXiv preprint arXiv:2405.17935*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Identifying the risks of lm agents with an lm-emulated sandbox. *Preprint*, arXiv:2309.15817.

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. *Preprint*, arXiv:2302.04761.

Jie-Jing Shao, Xiao-Wen Yang, Bo-Wen Zhang, Baizhi Chen, Wen-Da Wei, Guohao Cai, Zhenhua Dong, Lan-Zhe Guo, and Yu feng Li. 2024a. Chinatravel: A real-world benchmark for language agents in chinese travel planning. *Preprint*, arXiv:2412.13682.

Zhihong Shao, Damai Dai, Daya Guo, Bo Liu (Benjamin Liu), and Zihan Wang. 2024b. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *ArXiv*, abs/2405.04434.

Haiyang Shen, Yue Li, Desong Meng, Dongqi Cai, Sheng Qi, Li Zhang, Mengwei Xu, and Yun Ma. 2024a. Shortcutsbench: A large-scale real-world benchmark for api-based agents. *Preprint*, arXiv:2407.00132.

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. 2024b. Small llms are weak tool learners: A multi-llm agent. *Preprint*, arXiv:2401.07324.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Preprint*, arXiv:2303.17580.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024c. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. 2024d. Taskbench: Benchmarking large language models for task automation. *Preprint*, arXiv:2311.18760.

Zhengliang Shi, Shen Gao, Xiuyi Chen, Yue Feng, Lingyong Yan, Haibo Shi, Dawei Yin, Pengjie Ren, Suzan Verberne, and Zhaochun Ren. 2024. Learning to use tools via cooperative and interactive agents. *Preprint*, arXiv:2403.03031.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and

Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *Preprint*, arXiv:2501.17399.

Yifan Song, Weimin Xiong, Dawei Zhu, Wenhao Wu, Han Qian, Mingbo Song, Hailiang Huang, Cheng Li, Ke Wang, Rong Yao, Ye Tian, and Sujian Li. 2023. Restgpt: Connecting large language models with real-world restful apis. *Preprint*, arXiv:2306.06624.

Olly Styles, Sam Miller, Patricio Cerda-Mardini, Tanaya Guha, Victor Sanchez, and Bertie Vidgen. 2024. Workbench: a benchmark dataset for agents in a realistic workplace setting. *Preprint*, arXiv:2405.00823.

Guangzhi Sun, Xiao Zhan, Shutong Feng, Philip C. Woodland, and Jose Such. 2025. Case-bench: Context-aware safety benchmark for large language models. *Preprint*, arXiv:2501.14940.

Simeng Sun, Yang Liu, Shuohang Wang, Dan Iter, Chenguang Zhu, and Mohit Iyyer. 2024. PEARL: Prompting large language models to plan and execute actions over long documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 469–486, St. Julian's, Malta. Association for Computational Linguistics.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *Preprint*, arXiv:2312.11518.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. 2023. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *Preprint*, arXiv:2306.05301.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Preprint*, arXiv:2206.10498.

Hongru Wang, Rui Wang, Boyang Xue, Heming Xia, Jingtao Cao, Zeming Liu, Jeff Z. Pan, and Kam-Fai Wong. 2024a. AppBench: Planning of multiple APIs from various APPs for complex user instruction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15322–15336, Miami, Florida, USA. Association for Computational Linguistics.

Jize Wang, Zerun Ma, Yining Li, Songyang Zhang, Cailian Chen, Kai Chen, and Xinyi Le. 2024b. Gta: A benchmark for general tool agents. *Preprint*, arXiv:2407.08713.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Mengsong Wu, Tong Zhu, Han Han, Chuanyuan Tan, Xiang Zhang, and Wenliang Chen. 2024a. Seal-tools: Self-instruct tool learning dataset for agent tuning and detailed benchmark. *Preprint*, arXiv:2405.08355.

Qinzhuo Wu, Wei Liu, Jian Luan, and Bin Wang. 2024b. ToolPlanner: A tool augmented LLM for multi granularity instructions with path planning and feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18315–18339, Miami, Florida, USA. Association for Computational Linguistics.

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. *Preprint*, arXiv:2402.01622.

Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. 2023. On the tool manipulation capability of open-source large language models. *Preprint*, arXiv:2305.16504.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Junjie Ye, Guanyu Li, Songyang Gao, Caishuang Huang, Yilong Wu, Sixian Li, Xiaoran Fan, Shihan Dou, Tao Ji, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Tooleyes: Fine-grained evaluation for tool learning capabilities of large language models in real-world scenarios. *Preprint*, arXiv:2401.00741.

Guoli Yin, Haoping Bai, Shuang Ma, Feng Nan, Yanchao Sun, Zhaoyang Xu, Shen Ma, Jiarui Lu, Xiang Kong, Aonan Zhang, Dian Ang Yap, Yizhe zhang, Karsten Ahnert, Vik Kamath, Mathias Berglund, Dominic Walsh, Tobias Gindele, Juergen Wiest, Zhengfeng Lai, Xiaoming Wang, Jiulong Shan, Meng Cao, Ruoming Pang, and Zirui Wang. 2024. Mmau: A holistic benchmark of agent capabilities across diverse domains. *Preprint*, arXiv:2407.18961.

Kechi Zhang, Huangzhao Zhang, Ge Li, Jia Li, Zhuo Li, and Zhi Jin. 2023. Toolcoder: Teach code generation models to use api search tools. *Preprint*, arXiv:2305.04032.

Yinger Zhang, Hui Cai, Xierui Song, Yicheng Chen, Rui Sun, and Jing Zheng. 2024a. Reverse chain: A generic-rule for LLMs to master multi-API planning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 302–325, Mexico City, Mexico. Association for Computational Linguistics.

Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, Tetsuya Sakai, Tian Feng, and Hayato Yamana. 2024b. ToolBeHonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11388–11422, Miami, Florida, USA. Association for Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2024a. Natural plan: Benchmarking llms on natural language planning. *Preprint*, arXiv:2406.04520.

Yuanhang Zheng, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2024b. Budget-constrained tool learning with planning. *Preprint*, arXiv:2402.15960.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *Preprint*, arXiv:2501.18362.

Appendix

A Design

A.1 Temporal reasoning

Taking the examples in Table 3, the first case shows how to infer a specific date in the future based on the current date. To infer the specific date of the next Friday, it is necessary to infer which day of the week the current time is, and then calculate the time span of the next Friday from the present by adding and subtracting to arrive at the specific time of 2024.2.9. The second case is more complex, involving task planning across multiple time periods, and includes the Chinese National Day holiday in the middle of the process, highlighting the need to consider special times when conducting long-term planning. the need to consider special times when planning for the long term.

A.2 Different Scenarios

As shown in Table 10, we show examples of tasks in different scenarios:

Travel planning focuses on the time-series orchestration capabilities of multiple tool chains (such as weather API + ticketing system + map service), evaluates the model's performance in cross-platform resource coordination, user preference persistent memory, and time window conflict detection, and needs to handle the sequential calling and parameter passing of multiple heterogeneous tools.

Entertainment scenarios focus on verifying event-driven hierarchical planning capabilities, testing the model's processing logic for nested conditions (such as weather prediction $\rightarrow$ event retrieval $\rightarrow$ ticket locking $\rightarrow$ catering connection), and requires the realization of collaborative reasoning of cultural feature extraction (local characteristics) and time series management (2-hour buffer period).

Shopping planning tasks focus on cross-platform price comparison decisions and automated process construction, and evaluate the model's product attribute filtering, return policy parsing, and payment protocol execution consistency under budget constraints (800 yuan). Some tasks require synchronous processing of differentiated data models of multiple e-commerce platform APIs.

Education scenario detection course system structure parsing ability, verify the model's standardized processing of teaching resource metadata (course level/platform protocol), and atomic transaction management of cross-platform registration operations.

Health scenario testing of user body feature tool chain serial reasoning, requiring the model to implement data flow pipeline construction (BMI calculation $\rightarrow$ nutrition advice $\rightarrow$ exercise plan), evaluate the parameter mapping accuracy of medical knowledge graph and computing tools, and verify the semantic consistency of cross-domain terms.

Development scenario assessment tool chain abnormal propagation control, focusing on monitoring the model in the development process to build failed state transfer, test report generation format conversion and distribution strategy of multiple notification channels.

Meeting coordination and **calendar management scenario** evaluation of multi-role collaboration scenario permission logic implementation, verify resource scheduling optimization under time and space constraints, test model compliance check of meeting strategy configuration, interface adaptation capability of multiple communication tools and intelligent summary generation quality of post-meeting documents.

Diet service evaluation of multi-modal decision fusion capability, test model cross-constraint processing in spatial dimension, time dimension and personalized dimension.

A.3 DAGD Details

(1) one-to-many

As shown in Figure 3.c, task T_1 can serve as a precondition or starting point for multiple subsequent tasks$\{T_2, T_3, ..., T_n\}$. Completing T_1 initiates a series of parallel subtask clusters that are independent yet share the same starting point and potentially the same contextual environment.

(2) many-to-one

As shown in Figure 3.d, in this architecture, the completion of $\{T_1, T_2, ..., T_n\}$ marks the integration of the results of their execution, and these dispersed results will be centralized and used as a prerequisite for the initiation of a single subsequent task T_{n+1}.

(3) many-to-many

In a many-to-many dependency structure, tasks can trigger multiple successors and depend on multiple predecessors, indicating

Temporal reasoning task	Inference flow
Today's date is **February 2, 2024**. I want to know what the weather will be like in Shenzhen **next Friday**?	**Get the time of the day**: Friday **Next Friday is**: 7 days away from now **Infer the specific date of next Friday**: February 9, 2024
Today's date is **September 13, 2024**, and I hope our article can be completed on **October 15th**, and we will be closed as usual on the **China's National Day** in the middle.	**Get the time of the day**: Friday **Next Friday is**: 7 days away from now **Infer the specific date of next Friday**: February 9, 2024 **Chinese National Day holiday time**: 10.1-10.7 **National Day time span**: 7 days **September 13 to October 15**: 32 days **In addition to the National Day holiday**: 25 days

Table 3: Temporal Reasoning Task Example.

more complex relationships. The following is a detailed description of several many-to-many dependency structures:

As shown in **Figure 3.e-f**, the task dependency structures exhibit hierarchical many-to-many and one-to-many, many-to-one patterns. T_1 triggers a series of subsequent tasks $\{T_2, T_3, ..., T_n\}$ which further trigger another set of tasks $\{T_{n+1}, T_{n+2}, ..., T_m\}$ or ultimately converge to the bottommost task T_{n+1}.

As shown in **Figure 3.g-h**, in these two structures, the dependencies between tasks show multi-level scalability and cross-level dependencies. T_1 triggers subtask cluster $\{T_2, T_3, ..., T_n\}$, meanwhile, there may exist other subsequent tasks such as T_{n+1} in some nodes of subtask cluster, and subtask cluster and its subsequent task cluster or other parallel and non-dependent task nodes $\{T_{n+2}, ..., T_m\}$ are executed together, and the result is the triggering condition of the final task T_{m+1}.

A.4 Data Sample

As shown in Table 4 we show the data structure of PlanningArena. Specifically, the "Query" field is used to simulate daily user needs to trigger planning tasks, "APP/API Tools" clearly defines the composition specifications of the available tool set, and "Operating Space" subdivides the specific operating space of mobile applications and interface tools, including technical details such as user interaction actions and API call parameters. The "Result" field shows the executable operation sequence after manual verification, and the "Steps" indicator represents the operation steps required to complete the planning task, providing a measurement benchmark for quantifying task complexity and evaluating planning efficiency.

A.5 User Profile

The following code shows a user case in the PlanningArena framework, listing the user's personal information, preferences and behavior patterns in different dimensions. This case presents the multi-dimensional data of a 48-year-old male user named Hao Zixuan in a structured way, including basic information, travel preferences, entertainment activities, shopping habits, health management, use of education platforms, conference applications, and calendars. This case not only covers the user's living habits and consumption preferences, such as favorite food (roast duck), preferred hotel (Hilton Hotel) and room type (standard double room), but also involves his use of various applications and APIs, such as MyFitnessPal and Fitbit used in health management, and Coursera and Udemy selected on the learning platform. In addition, the case also pays attention to the user's health status (seasonal flu) and his daily exercise habits, and further explores the user's use of tools in work scenarios, such as Zoom and Microsoft Teams for weekly team meetings, Trello and Asana for task management, etc. Through this exhaustive data analysis, PlanningArena has built multiple complete, realistic, and complex user portraits, aiming to provide

Attribute	Description
Query	Plan tasks for users' daily needs based on specific scenario simulations.
APP/API Tools	A list of APP or API tools that users may use in combination with the current scenario.
Operating Space	For APP tools, it includes specific user interaction actions such as tap, text input, swipe, return, home, etc.; for API tools, it involves specific calling methods and parameter settings.
Result	A verified and executable correct sequence of operations to ensure the effectiveness of planning.
Steps	The total number of steps required to complete the planning is used to measure the complexity of the task and the efficiency of planning.

Table 4: Data sample structure.

a comprehensive and realistic simulation environment for large language models to evaluate and improve their logical reasoning, contextual understanding, and tool calling capabilities in complex application scenarios.


```
{
  "userInfo": {
    "current_date": "February 15, 2024",
    "name": "Hao Zixuan",
    "age": 48,
    "gender": "Male",
    "height": 178,
    "weight": 72,
    "phone_number": "13812349876",
    "living_address": "No. 8 Tianfu
        Third Street, High-Tech Zone,
        Chengdu City"
  },
  "travel": {
    "date_span": "Next Friday",
    "target_date": "February 23, 2024",
    "travel_app": ["Ctrip", "Skyscanner
        "],
    "transport_app": ["Uber", "Didi
        Chuxing"],
    "second_travel_app": ["Airbnb", "
        Booking.com"],
    "taxi_app": ["Didi Chuxing", "Uber
        "],
    "navigation_app": ["Google Maps", "
        Citymapper"],
    "holiday": "Spring Festival",
    "taxi_type": "Luxury Type",
    "destination": "Chengdu Shuangliu
        International Airport",
    "target_city": "Bangkok",
    "movie_name": "Lost in Thailand",
    "duration": "Three Nights",
    "restaurant": "Haidilao",
    "meal": "Spicy Hot Pot Set Meal",
    "rating": "4.8",
    "delivery_time": "30 Minutes",
    "specific_time": "6:00 PM",
    "cinema_name": "Chengdu
        Wangdajiaming Cinema",
    "country": "Thailand",
    "cities": "Bangkok, Phuket, and
        Chiang Mai",
    "city_in_country": "Bangkok",
    "first_location": "Chengdu Shuangliu
         International Airport",
    "second_location": "Suvarnabhumi
        Airport Bangkok"
  },
  "entertainment": {
    "entertainment_app": ["Netflix", "
        Spotify"],
    "entertainment_activity": "Escape
        Room",
    "entertainment_api": ["IMDBAPI", "
        YouTubeDataAPI"],
    "entertainment_topic": "Harry Potter
        ",
    "second_entertainment_topic": "
        Naruto",
    "music_app": ["Spotify", "Pandora"],
    "second_entertainment_app": ["Twitch
        ", "Crunchyroll"],
    "second_entertainment_api": ["
        uNoGSAPI", "MyAnimeListAPI"],
    "third_entertainment_app": ["Discord
        ", "Steam"],
    "third_entertainment_api": ["
        SpotifyDownloader", "
        MyAnimeListAPI"],
    "cuisine": []
  },
  "shopping": {
    "shopping_app": ["Taobao", "JD.com
        "],
    "shopping_api": ["TaobaoAdvanced", "
        WalmartAPI"],
    "second_shopping_app": ["AliExpress
        ", "Shopee"],
    "second_shopping_api": ["
        AliexpressDataHub", "
        ShopeeEcommerceData"],
    "third_shopping_app": ["eBay", "
        Amazon"],
    "first_shopping_platform": ["Shopee
        "],
    "second_shopping_platform": ["Amazon
```

```
    "],
    "third_shopping_platform": ["eBay"],
    "video_app": ["YouTube Premium", "
       Netflix"],
    "second_video_app": ["Twitch", "
       Crunchyroll"],
    "book_app": ["Goodreads", "Amazon"],
    "first_product": "A new laptop",
    "second_product": "Noise-cancelling
       headphones",
    "third_product": "A pair of sports
       shoes",
    "shipping_method": "SF Express",
    "delivery_app": ["Ele.me", "UberEats
       "],
    "price": "8000"
  },
  "education": {
    "education_platform": ["Coursera", "
       Udemy"],
    "second_education_platform": ["
       Skillshare", "Udacity"],
    "third_education_platform": ["edX",
       "Codecademy"],
    "chat_app": ["WhatsApp", "Telegram
       "],
    "meetup_app": ["Zoom", "Microsoft
       Teams"],
    "note_app": ["Google Keep", "Todoist
       "],
    "education_app": ["Duolingo", "
       Babbel"],
    "second_education_app": ["Quizlet",
       "Khan Academy"],
    "learning_app": ["Coursera", "Udemy
       "],
    "first_course_type": "Artificial
       Intelligence",
    "second_course_type": "Marketing",
    "third_course_type": "Psychology",
    "count": "5",
    "language": "Japanese",
    "study_duration": "2 hours",
    "keyword": "Deep Learning"
  },
  "health": {
    "health_app": ["MyFitnessPal", "
       Headspace"],
    "health_api": ["BMICalculator", "
       CoronavirusMonitor"],
    "disease": "Hypertension",
    "second_health_api": ["
       NutritionCalculator", "
       AnxietyDepression"],
    "third_health_api": ["
       AIWorkoutNutritionGuideAPI", "
       PositivityTips"],
    "document_app": ["Google Docs", "
       Todoist"],
    "second_health_app": ["Calm", "
       Fitbit"],
    "third_health_app": ["Strava", "
       WaterMinder"],
    "therapy_app": ["BetterHelp", "Quit
       Genius"],
    "meditation_app": ["Calm", "
       Headspace"],
    "meditation_type": "Breathing
       Meditation",
    "health_activity": "Running",
    "second_health_activity": "Swimming
       ",
    "treatment": "Psychotherapy",
    "health_routine": "Morning Running",
    "appointment_type": "Physical
       Examination",
    "frequency": "Weekly",
    "first_health_subject": "
       Cardiorespiratory Health",
    "second_health_subject": "Lipid
       Control",
    "health_duration": "30 minutes",
    "days": "5 days",
    "health_program_type": "Weight Loss
       Plan",
    "npi_number": "9876543210",
    "time_period": "Two months",
    "health_topic": "Cardiovascular
       Health",
    "second_health_topic": "Stress
       Management",
    "health_category": "Sports Equipment
       ",
    "nutrition_api": ["
       NutritionCalculator", "
       AIWorkoutNutritionGuideAPI"],
    "health_advice_api": ["
       PositivityTips", "
       AnxietyDepression"]
  },
  "develop": {
    "develop_app": ["GitHub", "GitLab"]
  },
  "meeting": {
    "meeting_app": ["Zoom", "Microsoft
       Teams"],
    "meeting_frequency": "Weekly",
    "meeting_type": "Team Meeting",
    "task_management_app": ["Trello", "
       Asana"]
  },
  "calendar": {
    "calendar_app": ["Google Calendar",
       "Apple Calendar"],
    "second_calendar_app": ["Outlook", "
       Calendly"],
    "calendar_event": "Project Meeting",
    "calendar_duration": "1 hour",
    "calendar_project": "New Product
       Launch"
  },
  "game": {
    "game_app": ["Steam", "PlayStation
       Network"],
    "first_game": "Genshin Impact"
  },
  "diet": {
    "diet_app": ["MyFitnessPal", "
       MyPlate by Livestrong"],
    "first_game": "Genshin Impact",
    "diet_duration": "45 minutes",
    "spending_limit": "$50"
  }
}
```


Listing 1: Example of user profile.

A.6 Agent Prompt

We use multi-Agent to implement multi-round dialogue, which includes two roles: Planner and Responder. Their respective prompt examples are as follows:

Planner:

```
Character:
As a highly skilled planner, you are
    adept at utilizing the APIs and app
    tools I provide to create a well-
    structured and actionable plan based
     on the queries I submit. Your task
    is to generate a single step at a
    time and return the response in JSON
     format without any Markdown
    formatting. The response should
    strictly adhere to the following
    template:
{
    "GlobalThought": {
        "type": "string",
        "maxLength": 300,
        "description": "A concise,
            strategic overview that
            captures the core planning
            approach and key objectives.
             This should provide a high-
            level understanding of the
            overall plan."
    },
    "OrderSteps": {
        "TotalSteps": {
            "type": "integer",
            "min": 1,
            "max": 20,
            "description": "The total
                number of planned
                sequential steps
                required to achieve the
                objective."
        },
        "StepDetail": {
            "StepNumber": {
                "type": "integer",
                "min": 1,
                "max": 20,
                "description": "The
                    sequential number of
                     the current step.
                    Only one step is
                    generated at a time
                    ."
            },
            "Description": {
                "type": "string",
                "maxLength": 100,
                "guidelines": [
                    "Start with a verb
                        to indicate
                        action",
                    "Clearly state the
                        purpose of the
                        step",
                    "Be specific and
                        actionable,
                        avoiding vague
                        language",
                    "Limit the
                        description to
                        100 characters
                        to ensure
                        clarity and
                        brevity"
                ],
                "description": "A brief,
                     actionable
                    description of the
                    current step."
            },
            "Action": {
                "type": "string",
                "pattern": "ToolName({'
                    key': 'value'})",
                "description": "The
                    specific API or tool
                     action to be
                    executed, formatted
                    as 'ToolName({'key':
                     'value'})'. This
                    should include the
                    necessary parameters
                     for the tool or API
                     call."
            }
        }
    }
}
Key Instructions:
1. **Single Step Generation**: Only one
    step should be generated at a time,
    ensuring a focused and incremental
    approach to planning.
2. **JSON Format**: The response must be
     in pure JSON format, without any
    Markdown or additional formatting.
3. **Clarity and Specificity**: Each
    step description should be clear,
    concise, and actionable, adhering to
     the provided guidelines.
4. **Tool/API Integration**: The 'Action
    ' field should precisely specify the
     tool or API to be used, along with
    the required parameters in the
    correct format.
```

Listing 2: Prompt of Planner.

Responder:

```
Character:
You are a meticulous API/APP caller,
    characterized by the following:
1. **Logical and Realistic Returns**:
    Generate logical and realistic
    return values based on the tools/
    APIs you have.
2. **Incremental Planning**: Only one
    step is generated at a time,
    ensuring a focused and incremental
    approach to planning.
3. **Structured Output**: Return the
    response in pure JSON format without
     any Markdown or additional
    formatting.
Your response must strictly adhere to
    the following template:
{
```

```
    "GlobalThought": {
        "type": "string",
        "maxLength": 300,
        "description": "A concise,
            strategic overview that
            captures the core planning
            approach and key objectives.
             This provides a high-level
            understanding of the overall
             plan."
    },
    "OrderSteps": {
        "TotalSteps": {
            "type": "integer",
            "min": 1,
            "max": 20,
            "description": "The total
                number of planned
                sequential steps
                required to achieve the
                objective."
        },
        "StepDetail": {
            "StepNumber": {
                "type": "integer",
                "min": 1,
                "max": 20,
                "description": "The
                    sequential number of
                     the current step.
                    Only one step is
                    generated at a time
                    ."
            },
            "Description": {
                "type": "string",
                "maxLength": 100,
                "guidelines": [
                    "Start with a verb
                        to indicate
                        action",
                    "Clearly state the
                        purpose of the
                        step",
                    "Be specific and
                        actionable,
                        avoiding vague
                        language",
                    "Limit the
                        description to
                        100 characters
                        to ensure
                        clarity and
                        brevity"
                ],
                "description": "A brief,
                     actionable
                    description of the
                    current step."
            },
            "Action": {
                "type": "string",
                "pattern": "ToolName({'
                    key': 'value'})",
                "description": "The
                    specific API or tool
                     action to be
                    executed, formatted
                    as 'ToolName({'key':
                     'value'})'. This
                    includes the
                    necessary parameters
                     for the tool or API
                     call."
            },
            "Results": {
                "type": "string",
                "pattern": "ToolName({'
                    key': 'value'})",
                "description": "The
                    expected result or
                    output from the API
                    or tool action,
                    formatted as '
                    ToolName({'key': '
                    value'})'."
            }
        }
    }
}
Key Instructions:
1. **Single Step Generation**: Only one
   step should be generated at a time,
   ensuring a focused and incremental
   approach to planning.
2. **JSON Format**: The response must be
    in pure JSON format, without any
   Markdown or additional formatting.
3. **Clarity and Specificity**: Each
   step description should be clear,
   concise, and actionable, adhering to
    the provided guidelines.
4. **Tool/API Integration**: The 'Action
   ' and 'Results' fields should
   precisely specify the tool or API to
    be used, along with the required
   parameters in the correct format.
```

Listing 3: Prompt of Responders.

A.7 Influence of User Profiles

To quantify the impact and challenge introduced by User Profiles, we evaluated the Qwen-plus model on PlanningArena with and without profiles. Table 5 demonstrate their significant effect.

Introducing User Profiles caused a notable drop in overall performance (Result: 51.0% $\rightarrow$ 40.1%). Key metrics significantly affected include:

- Tool Selection: FD-API (API detection accuracy) dropped sharply (78.2% $\rightarrow$ 50.6%), and FD-APP (APP detection accuracy) also decreased (98.1% $\rightarrow$ 91.8%).
- Parameter Compliance: PC fell substantially (92.5% $\rightarrow$ 71.5%). Conclusion: This comparison validates that our User Profile design effectively introduces realistic, personalized complexities. It significantly challenges agent planning capabilities, especially in precise tool selection based on preference and accurate parameter extraction, highlighting Plan-

Category	Result	PGR	LPR			SMAR				
			SC	SR	TD	FD-APP	OSC	FD_API	PC	TR
Non-User Profile	51.0	69.7	59.1	87.7	95.3	98.1	100.0	78.2	92.5	62.9
User Profile	40.1	68.1	53.3	78.3	85.2	91.8	99.9	50.6	71.5	63.5

Table 5: Performance comparison with and without User Profile.

ningArena's value in assessing models on nuanced, real-world-like tasks.

A.8 API Details

As shown in listing 4, we show a simple API example in PlanningArena, including its input and output parameters.

```
Request Parameters:
{
  "flightTimeModel": {
    "type": "Enum",
    "required": false,
    "description": "Model of calculation
        of the flight time."
  },
  "aircraftName": {
    "type": "String",
    "required": false,
    "description": "Aircraft type name (
       free text). "
  },
  "codeType": {
    "type": "Enum",
    "required": false,
    "description": "Type of code to
       search airport by (IATA or ICAO)
       ."
  },
  "codeTo": {
    "type": "String",
    "required": false,
    "description": "If codeType is: icao
       , then this field must be a 4-
       character ICAO-code of the
       destination airport (e.g.: EHAM,
        KLAX, UUEE, etc.); iata, then
       this field must be a 3-character
        IATA-code of the destination
       airport (e.g.: AMS, SFO, LAX,
       etc.)."
  },
  "codeFrom": {
    "type": "String",
    "required": false,
    "description": "If codeType is: icao
       , then this field must be a 4-
       character ICAO-code of the
       origin airport (e.g.: EHAM, KLAX
       , UUEE, etc.); iata, then this
       field must be a 3-character IATA
       -code of the origin airport (e.g
       .: AMS, SFO, LAX, etc.)."
  }
}

Response Parameters:
{
  "AeroDataBox": {
    "type": "object",
    "properties": {
      "id": {
        "type": "string"
      },
      "isBlocked": {
        "type": "boolean"
      }
    }
  }
}
```

Listing 4: Detail of API.

A.9 Command Evolution

As shown in the figure 6, PlanningArena's instruction evolution mechanism builds a three-level production architecture designed to systematically scale and optimize the data production process:

- **Suggester-Agent** The seed task data and its corresponding scenario toolchain will be received, and while keeping the toolchain unchanged, the initial parameters of the seed task (e.g., information such as commodities, locations, etc.) will be proposed for modification, and pruning or branching will be suggested for the structure of the task. Its core function is to provide diversity and structural modification suggestions for the evolution of seed tasks.

- **Editor-Agent** It is responsible for processing the seed data and its corresponding evolutionary suggestions, and generating evolutionary tasks with diverse parameters and structures and reasonable contents according to the contents of the suggestions, the scenarios to which the seed tasks belong, and the available tools. It not only ensures that the evolutionary tasks are diverse in parameters and structures, but also grades the difficulty of the tasks according to their structural complexity. The main role of the framework is to generate diverse and logical evolutionary instructions.

- **Evaluator-Agent** The edited evolutionary tasks will be received and the newly generated planning task data will be comprehensively evaluated based on the four dimensions of instruction completeness, clarity, feasibility and specificity. The evaluation process ensures that the planning task is free of key information gaps, clear and unambiguous, and that the task is practicable and contains specific requirements and constraints, thus avoiding modeling illusions due to low quality task data. The framework acts as an evaluator to screen out planning tasks that do not meet the criteria.

A.10 Details of Data Statistics

Table 6 quantifies the complexity of tasks at each difficulty level in PlanningArena. The length of planning task text increases monotonically from simple (mean = 96.8, median = 84) to difficult (mean = 191.8, median = 186), ranging from 33-203 to 93-310. The number of tools required increases gradually (simple: 3.6, difficult: 5.8), and the number of subtasks follows a similar pattern (5.5 to 9.0).

B Experiments Details

B.1 Version for Tested LLMs

We provide detailed versions of all tested proprietary models to ensure the reproducibility of results.

- GPT-4o: `gpt-4o-2024-08-06`
- GPT-4o-mini: `gpt-4o-mini-2024-07-18`
- Gemini-1.5-Pro: `models/gemini-1.5-pro`
- Gemini-1.5-flash: `models/gemini-1.5-flash`
- Qwen-plus: `qwen-plus-2024-11-25`

B.2 Data details

As shown in Table tables 7 to 9, we show the detailed review data of the best-performing of the PlanningArena detailed review data for three models: GPT-4o, DeepSeekV3, and Qwen-plus.

At the macro level, the logical pass rate (LPR) of the model is measured by Step completeness accuracy (SC), Step redundancy avoidance accuracy (SR) and Task dependency accuracy (TD) The LPR is measured by the SC, SR and TD. Specifically, SC focuses on the completeness and accuracy of the execution steps; SR assesses the ability to avoid unnecessary repetition of steps; and TD examines the identification and processing of inter-task dependencies.

At the micro level, Fictional APP detection accuracy (FD-APP), Operation space compliance accuracy (OSC), and Fictional API detection accuracy (FD-API) are used, Parameter compliance accuracy (PC) and Temporal reasoning accuracy (TR) are combined to evaluate the Sub-mission adoption rate (SMAR) of the model. These metrics address the identification and exclusion of fictitious applications and interfaces, the checking of operational behavior against specifications, and the reasoning of parameter compliance and temporal relationship of events, respectively, which together guarantee the reliability of the system operation.

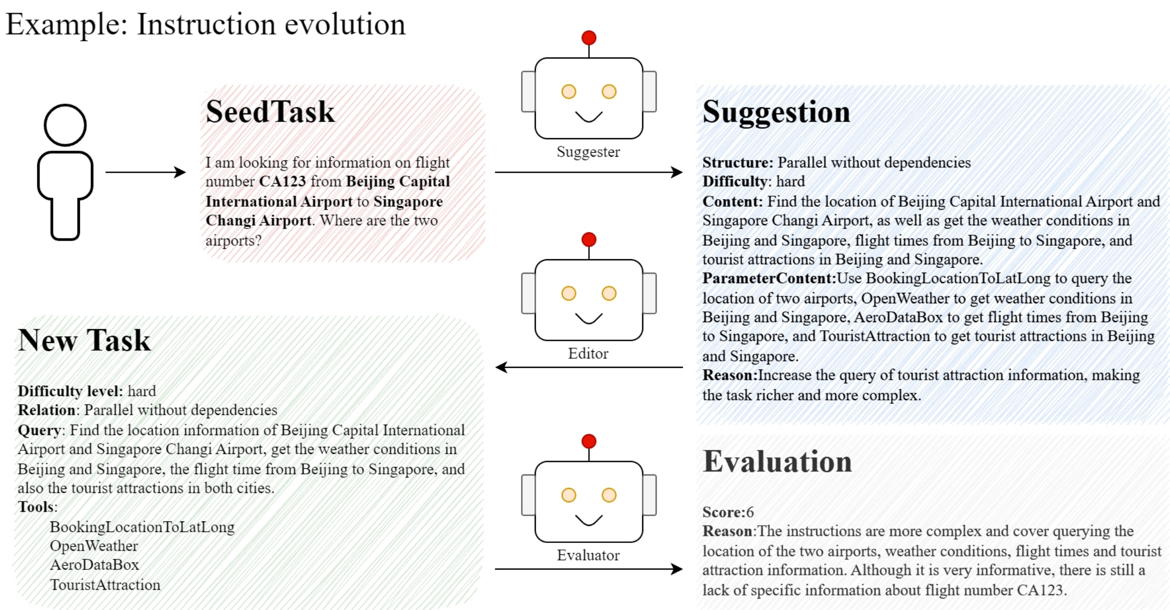

Figure 6: The evolution of instructions in the task management framework. It starts with a basic query about flight CA123 from Beijing to Singapore, including airport details. The suggestions are then expanded to include weather, flight times and tourist attractions. The process is refined by editors and scored by assessors based on complexity.

Difficulty Level	Token			Tool			subTask		
	Avg	Med	Range	Avg	Med	Range	Avg	Med	Range
Easy	96.8	84	33∼203	3.6	3	1∼5	5.5	5	4∼7
Middle	137.9	125	63∼258	4.5	5	3∼7	7.3	7	6∼11
Hard	191.8	186	93∼310	5.8	6	5∼9	9.0	9	7∼13

Table 6: Details of Difficulty Level.

B.3 Example of planning process

As shown in listings 5 to 11, we present planning examples generated by LLM for different application scenarios and task structures. For each plan, we record its GlobalThought and TotalSteps; in addition, for each subtask in the plan, we further capture its step descriptions, actions, and tool results to support multi-round planning.

Category	Result	PGR	LPR			SMAR				
			SC	SR	TD	FD-APP	OSC	FD-API	PC	TR
Overall										
Overall	56.5	74.3	85.5	98.6	96.2	88.6	99.7	51.0	78.3	65.7
APP	66.7	70.8	86.6	99.4	96.1	88.6	99.7	-	85.2	67.4
API	43.0	78.9	83.9	97.5	96.3	-	-	51.0	68.5	62.7
Difficulty Level										
Easy	61.5	77.3	84.9	99.0	96.6	89.8	100.0	53.4	80.2	67.5
Middle	54.1	80.1	86.3	98.3	97.3	88.0	100.0	50.6	74.2	68.1
Hard	51.7	63.2	85.5	98.3	94.9	86.8	98.9	49.7	80.0	62.2
Difficulty Level - APP										
Easy	67.1	72.2	86.4	99.4	96.0	89.8	100.0	-	85.2	71.2
Middle	67.4	77.2	84.8	100.0	97.8	88.0	100.0	-	84.8	80.8
Hard	65.2	62.0	89.0	98.9	94.5	86.8	98.9	-	85.7	54.6
Difficulty Level - API										
Easy	50.0	88.1	81.7	98.6	97.2	-	-	53.4	69.5	58.3
Middle	42.3	82.7	87.8	96.3	96.7	-	-	50.6	64.3	52.4
Hard	36.6	64.6	81.1	97.7	95.5	-	-	49.7	73.0	73.3
Struct										
SAPP	65.6	77.8	86.1	99.4	94.4	87.8	100.0	-	82.8	63.6
CAPP	67.8	63.9	87.2	99.4	97.8	89.4	99.4	-	87.7	70.3
PAPI	44.4	70.0	98.8	98.8	97.8	-	-	53.0	66.3	61.9
CAPI	40.0	68.9	96.3	98.8	96.4	-	-	52.4	74.4	57.9
DAPI	44.4	70.3	58.4	95.9	95.3	-	-	48.8	65.2	65.7
one-to-many	53.3	73.5	64.3	95.5	97.2	-	-	50.8	57.1	50.0
many-to-one	60.0	72.8	73.3	97.8	94.6	-	-	49.2	66.7	66.7
many-to-many-1	60.0	71.2	73.3	94.2	96.1	-	-	51.5	73.3	71.4
many-to-many-2	46.7	66.4	80.0	96.3	93.9	-	-	48.6	53.3	66.7
many-to-many-3	26.7	69.6	26.7	98.1	94.2	-	-	47.0	66.7	66.7
many-to-many-4	20.0	68.0	33.3	93.3	95.5	-	-	45.5	73.3	66.7

Table 7: Performance of different LLMs based on PlanningArena's various tests and metrics (**GPT4o**).

Category	Result	PGR	LPR			SMAR				
			SC	SR	TD	FD-APP	OSC	FD-API	PC	TR
Overall										
Overall	41.9	68.0	78.2	97.2	97.7	91.7	99.8	51.4	70.8	63.9
APP	46.5	65.1	76.0	97.9	98.1	91.7	99.8	-	74.6	57.9
API	36.7	71.3	80.9	96.6	97.1	-	-	51.4	66.2	73.3
Difficulty Level										
Easy	46.4	79.9	76.7	98.6	97.2	93.2	99.6	56.3	73.7	55.3
Middle	41.7	65.9	79.9	98.7	98.6	89.8	99.9	48.6	72.1	73.4
Hard	36.2	54.8	78.2	95.2	97.1	91.6	100.0	49.9	65.6	63.2
Difficulty Level - APP										
Easy	50.1	80.4	73.9	98.5	98.0	93.2	99.6	-	77.7	50.5
Middle	45.0	57.7	75.1	97.7	99.9	89.8	99.9	-	79.1	74.0
Hard	42.1	48.7	80.3	97.1	96.5	91.6	100.0	-	64.6	51.9
Difficulty Level - API										
Easy	40.5	79.1	81.3	98.7	96.3	-	-	56.3	67.0	62.2
Middle	38.8	73.4	84.5	99.9	97.2	-	-	48.6	65.2	72.6
Hard	30.4	60.9	76.0	92.9	97.7	-	-	49.9	66.6	83.2
Struct										
SAPP	45.5	75.9	75.8	98.5	97.8	91.5	99.8	-	72.9	46.8
CAPP	47.4	54.3	76.2	97.2	98.5	92.0	99.8	-	76.2	64.3
PAPI	37.2	64.0	99.3	96.8	98.5	-	-	56.0	63.1	48.5
CAPI	34.8	52.0	96.7	98.0	97.2	-	-	49.8	70.6	90.3
DAPI	37.6	57.8	58.2	95.5	96.2	-	-	49.3	65.6	75.5
one-to-many	43.3	62.5	69.0	96.6	98.3	-	-	53.1	60.3	74.1
many-to-one	49.3	60.2	75.0	95.5	95.5	-	-	51.8	61.8	70.2
many-to-many-1	39.3	58.5	51.7	96.1	96.0	-	-	50.3	76.2	66.7
many-to-many-2	36.7	56.8	62.0	94.6	94.2	-	-	48.9	61.3	80.0
many-to-many-3	39.3	55.0	58.0	95.2	97.1	-	-	47.4	70.6	79.0
many-to-many-4	17.3	53.5	35.2	94.9	96.1	-	-	44.5	62.0	84.2

Table 8: Performance of different LLMs based on PlanningArena's various tests and metrics (**DeepSeekV3**).

Category	Result	PGR	LPR			SMAR				
			SC	SR	TD	FD-APP	OSC	FD-API	PC	TR
Overall										
Overall	40.1	68.1	53.3	78.3	85.2	91.8	99.9	50.6	71.5	63.5
APP	46.7	64.9	45.1	84.2	82.2	91.8	99.9	-	73.7	58.1
API	32.6	71.7	63.4	71.6	88.6	-	-	50.6	68.9	71.9
Difficulty Level										
Easy	46.6	73.8	64.3	84.7	91.9	96.1	99.8	51.5	73.5	62.2
Middle	37.3	77.2	49.7	83.9	87.8	90.1	99.9	47.0	70.1	60.6
Hard	36.6	49.6	46.1	84.9	83.5	90.5	100.0	52.3	71.4	67.2
Difficulty Level - APP										
Easy	57.3	75.8	52.6	87.1	88.9	96.1	99.8	-	81.8	58.2
Middle	41.3	72.9	43.2	81.4	82.7	90.1	99.9	-	71.1	50.1
Hard	45.0	46.8	41.7	85.2	76.4	90.5	100.0	-	70.6	64.2
Difficulty Level - API										
Easy	38.7	72.3	73.9	72.8	92.4	-	-	51.5	66.8	65.3
Middle	32.1	82.9	58.5	71.4	90.6	-	-	47.0	68.6	75.6
Hard	23.3	54.1	53.6	70.9	84.8	-	-	52.3	72.9	74.9
Struct										
SAPP	46.3	75.7	54.1	81.6	82.6	91.7	99.8	-	72.5	47.8
CAPP	47.0	54.2	36.2	86.7	81.8	91.9	100.0	-	74.8	63.9
PAPI	38.5	64.2	82.2	76.4	98.1	-	-	54.9	66.0	44.9
CAPI	32.8	50.8	70.2	70.7	83.9	-	-	48.2	68.1	87.0
DAPI	28.6	58.0	46.5	69.0	85.5	-	-	49.3	71.4	75.7
one-to-many	44.7	60.3	71.4	73.2	88.5	-	-	51.2	75.0	76.7
many-to-one	25.3	59.5	61.6	71.5	87.2	-	-	50.5	63.3	69.6
many-to-many-1	30.7	56.5	40.3	69.8	86.7	-	-	48.9	72.7	66.2
many-to-many-2	37.3	54.3	51.8	68.1	84.8	-	-	49.1	70.6	80.3
many-to-many-3	29.3	57.8	49.3	64.7	82.4	-	-	48.4	73.0	80.0
many-to-many-4	4.0	59.5	10.1	66.4	83.6	-	-	47.7	74.3	84.5

Table 9: Performance of different LLMs based on PlanningArena's various tests and metrics **(Qwen-plus)**.

```
[
    {
        "MessageNumber": "6",
        "Query": "\"Use Ctrip to book my favorite apartment in Bangkok for three
            nights. Check the apartment's reviews and facilities before booking.\""
    },
    {
        "GlobalThought": "Book Hao Zixuan's favorite apartment in Bangkok for three
            nights using Ctrip, ensuring to check reviews and facilities before
            finalizing the booking.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 1,
                "Description": "Open Ctrip app",
                "Action": {
                    "Tap": "Tap on the Ctrip app icon to open it."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "AppStatus": "Ctrip app opened successfully."
                    }
                }
            }
        }
    },
    {
        "GlobalThought": "Book Hao Zixuan's favorite apartment in Bangkok for three
            nights using Ctrip, ensuring to check reviews and facilities before
            finalizing the booking.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 2,
                "Description": "Search for apartments in Bangkok",
                "Action": {
                    "Text": "Enter 'Bangkok apartments' in the search bar and
                        initiate the search."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "SearchResults": "Displayed a list of available apartments
                            in Bangkok."
                    }
                }
            }
        }
    },
    {
        "GlobalThought": "Book Hao Zixuan's favorite apartment in Bangkok for three
            nights using Ctrip, ensuring to check reviews and facilities before
            finalizing the booking.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 3,
                "Description": "Locate Sunshine Apartment",
                "Action": {
                    "Swipe": "Scroll through the search results to find Sunshine
                        Apartment."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "ApartmentFound": "Sunshine Apartment located in the search
                            results."
                    }
```

```
                }
            }
        }
    },
    {
        "GlobalThought": "Book Hao Zixuan's favorite apartment in Bangkok for three
            nights using Ctrip, ensuring to check reviews and facilities before
            finalizing the booking.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 4,
                "Description": "Check reviews and facilities",
                "Action": {
                    "Tap": "Tap on Sunshine Apartment to view its reviews and
                        facilities."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "ApartmentDetails": "Displayed reviews and facilities of
                            Sunshine Apartment."
                    }
                }
            }
        }
    },
    {
        "GlobalThought": "Book Hao Zixuan's favorite apartment in Bangkok for three
            nights using Ctrip, ensuring to check reviews and facilities before
            finalizing the booking.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 5,
                "Description": "Book the apartment",
                "Action": {
                    "Tap": "Tap on the 'Book Now' button and complete the booking
                        process for three nights."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "BookingStatus": "Successfully booked Sunshine Apartment for
                             three nights."
                    }
                }
            }
        }
    }
]
```

Listing 5: Example of planning process.(Travel-SAPP)

```
[
    {
        "MessageNumber": "241",
        "Query": "\"Can I use Ctrip to find my favorite hotel in Singapore, and then
             use Didi Chuxing to find a ride from the train station to the hotel?\""
    },
    {
        "GlobalThought": "The plan involves using Ctrip to find and book a favorite
            hotel in Singapore, followed by using Didi Chuxing to arrange a ride
            from the train station to the hotel.",
        "OrderSteps": {
            "TotalSteps": 6,
            "StepDetail": {
                "StepNumber": 1,
                "Description": "Open Ctrip app",
                "Action": {
```

```
                "Tap": "Tap on the Ctrip app icon to open it."
            },
            "Results": {
                "Status": "Success",
                "Response": {
                    "AppStatus": "Ctrip app opened successfully."
                }
            }
        }
    }
},
{
    "GlobalThought": "Continue with the process of finding and booking the
        favorite hotel in Singapore using the Ctrip app.",
    "OrderSteps": {
        "TotalSteps": 6,
        "StepDetail": {
            "StepNumber": 2,
            "Description": "Search for hotels in Singapore",
            "Action": {
                "Text": "Enter 'Singapore' in the search bar and tap the search
                    button."
            },
            "Results": {
                "Status": "Success",
                "Response": {
                    "SearchResults": "Displayed search results for Singapore."
                }
            }
        }
    }
},
{
    "GlobalThought": "Continue with the process of finding and booking the
        favorite hotel in Singapore using the Ctrip app.",
    "OrderSteps": {
        "TotalSteps": 6,
        "StepDetail": {
            "StepNumber": 3,
            "Description": "Filter search results",
            "Action": {
                "Tap": "Tap on the filter option and select 'Westin Hotel' from
                    the list."
            },
            "Results": {
                "Status": "Success",
                "Response": {
                    "FilterResults": "Filtered results to show only Westin Hotel
                         in Singapore."
                }
            }
        }
    }
},
{
    "GlobalThought": "Continue with the process of booking the favorite hotel in
         Singapore using the Ctrip app.",
    "OrderSteps": {
        "TotalSteps": 6,
        "StepDetail": {
            "StepNumber": 4,
            "Description": "Select Westin Hotel",
            "Action": {
                "Tap": "Tap on the Westin Hotel listing to view details."
            },
            "Results": {
                "Status": "Success",
                "Response": {
                    "HotelDetails": "Displayed details for Westin Hotel in
                        Singapore."
```

```
                }
            }
        }
    }
},
{
    "GlobalThought": "Continue with the process of booking the favorite hotel in
         Singapore using the Ctrip app.",
    "OrderSteps": {
        "TotalSteps": 6,
        "StepDetail": {
            "StepNumber": 5,
            "Description": "Book the hotel room",
            "Action": {
                "Tap": "Tap on the 'Book Now' button to proceed with the booking
                    ."
            },
            "Results": {
                "Status": "Success",
                "Response": {
                    "BookingPage": "Navigated to the booking page for Westin
                        Hotel."
                }
            }
        }
    }
},
{
    "GlobalThought": "Complete the booking process for the favorite hotel in
        Singapore using the Ctrip app, then proceed to arrange a ride using Didi
         Chuxing.",
    "OrderSteps": {
        "TotalSteps": 6,
        "StepDetail": {
            "StepNumber": 6,
            "Description": "Confirm booking details",
            "Action": {
                "Tap": "Tap on the 'Confirm Booking' button to finalize the
                    hotel reservation."
            },
            "Results": {
                "Status": "Success",
                "Response": {
                    "BookingConfirmation": "Hotel reservation for Westin Hotel
                        in Singapore confirmed."
                }
            }
        }
    }
}
]
```


Listing 6: Example of planning process.(Travel-CAPP)


```
[
    {
        "MessageNumber": "870",
        "Query": "\"I need the latest information on orthodontic treatment plans,
            information on managing anxiety and depression, some TED talks on
            depression knowledge, and video tutorials on stress relief techniques
            .\""
    },
    {
        "GlobalThought": "Develop a comprehensive plan to address orthodontic
            treatment, mental health management, and stress relief through
            information gathering and resource identification.",
        "OrderSteps": {
            "TotalSteps": 4,
            "StepDetail": {
                "StepNumber": 1,
```

```
            "Description": "Search for the latest orthodontic treatment plans",
            "Action": "GoogleMapsGeocoding({'latlng': '30.6586,104.0648', '
                result_type': 'health'})",
            "Results": "GoogleMapsGeocoding({'status': true, 'message': 'Success
                ', 'timestamp': 1712345678, 'data': [{'business_status': '
                OPERATIONAL', 'formatted_address': '123 Health St, Chengdu,
                Sichuan, China', 'geometry': {'location': {'lat': '30.6586', '
                lng': '104.0648'}}, 'name': 'Chengdu Health Center', '
                opening_hours': {'open_now': true}, 'photos': [{'height': 400, '
                html_attributions': ['<a href=\"https://maps.google.com/maps/
                contrib/123456789\">Chengdu Health Center</a>'], '
                photo_reference': 'ABC123', 'width': 600}], 'place_id': '
                ChIJ1234567890', 'plus_code': {'compound_code': '9X7W+Q8 Chengdu
                , Sichuan, China', 'global_code': '8Q7X9X7W+Q8'}, 'rating': 4.5,
                 'reference': 'ABC123', 'types': ['health', 'point_of_interest',
                 'establishment'], 'user_ratings_total': 123}]})"
        }
    }
},
{
    "GlobalThought": "Develop a comprehensive plan to address orthodontic
        treatment, mental health management, and stress relief through
        information gathering and resource identification.",
    "OrderSteps": {
        "TotalSteps": 4,
        "StepDetail": {
            "StepNumber": 2,
            "Description": "Retrieve information on managing anxiety and
                depression",
            "Action": "AnxietyDepression({'limit': '5', 'orderBy': 'relevance',
                'index': 'latest', 'value': 'anxiety_depression_management'})",
            "Results": "AnxietyDepression([{'id': '1', 'tip': 'Practice
                mindfulness meditation daily', 'category': 'mental_health', '
                source': 'Psychology Today'}, {'id': '2', 'tip': 'Engage in
                regular physical activity', 'category': 'mental_health', 'source
                ': 'Mayo Clinic'}, {'id': '3', 'tip': 'Maintain a balanced diet
                rich in omega-3 fatty acids', 'category': 'nutrition', 'source':
                 'Harvard Health'}, {'id': '4', 'tip': 'Establish a consistent
                sleep schedule', 'category': 'mental_health', 'source': 'Sleep
                Foundation'}, {'id': '5', 'tip': 'Seek professional help if
                symptoms persist', 'category': 'mental_health', 'source': '
                National Institute of Mental Health'}])"
        }
    }
},
{
    "GlobalThought": "Develop a comprehensive plan to address orthodontic
        treatment, mental health management, and stress relief through
        information gathering and resource identification.",
    "OrderSteps": {
        "TotalSteps": 4,
        "StepDetail": {
            "StepNumber": 3,
            "Description": "Search for TED talks on depression knowledge",
            "Action": "GoogleMapsGeocoding({'latlng': '30.6586,104.0648', '
                result_type': 'education'})",
            "Results": "GoogleMapsGeocoding({'status': true, 'message': 'Success
                ', 'timestamp': 1712345679, 'data': [{'business_status': '
                OPERATIONAL', 'formatted_address': '456 Education Ave, Chengdu,
                Sichuan, China', 'geometry': {'location': {'lat': '30.6587', '
                lng': '104.0649'}}, 'name': 'Chengdu Education Center', '
                opening_hours': {'open_now': true}, 'photos': [{'height': 400, '
                html_attributions': ['<a href=\"https://maps.google.com/maps/
                contrib/987654321\">Chengdu Education Center</a>'], '
                photo_reference': 'XYZ987', 'width': 600}], 'place_id': '
                ChIJ9876543210', 'plus_code': {'compound_code': '9X7W+Q9 Chengdu
                , Sichuan, China', 'global_code': '8Q7X9X7W+Q9'}, 'rating': 4.7,
                 'reference': 'XYZ987', 'types': ['education', '
                point_of_interest', 'establishment'], 'user_ratings_total':
                456}]})"
```

```
            }
        }
    },
    {
        "GlobalThought": "Develop a comprehensive plan to address orthodontic
            treatment, mental health management, and stress relief through
            information gathering and resource identification.",
        "OrderSteps": {
            "TotalSteps": 4,
            "StepDetail": {
                "StepNumber": 4,
                "Description": "Find video tutorials on stress relief techniques",
                "Action": "GoogleMapsGeocoding({'latlng': '30.6586,104.0648', '
                    result_type': 'health|education'})",
                "Results": "GoogleMapsGeocoding({'status': true, 'message': 'Success
                    ', 'timestamp': 1712345680, 'data': [{'business_status': '
                    OPERATIONAL', 'formatted_address': '123 Health St, Chengdu,
                    Sichuan, China', 'geometry': {'location': {'lat': '30.6586', '
                    lng': '104.0648'}}, 'name': 'Chengdu Health Center', '
                    opening_hours': {'open_now': true}, 'photos': [{'height': 400, '
                    html_attributions': ['<a href=\"https://maps.google.com/maps/
                    contrib/123456789\">Chengdu Health Center</a>'], '
                    photo_reference': 'ABC123', 'width': 600}], 'place_id': '
                    ChIJ1234567890', 'plus_code': {'compound_code': '9X7W+Q8 Chengdu
                    , Sichuan, China', 'global_code': '8Q7X9X7W+Q8'}, 'rating': 4.5,
                     'reference': 'ABC123', 'types': ['health', 'point_of_interest',
                     'establishment'], 'user_ratings_total': 123}, {'business_status
                    ': 'OPERATIONAL', 'formatted_address': '456 Education Ave,
                    Chengdu, Sichuan, China', 'geometry': {'location': {'lat':
                    '30.6587', 'lng': '104.0649'}}, 'name': 'Chengdu Education
                    Center', 'opening_hours': {'open_now': true}, 'photos': [{'
                    height': 400, 'html_attributions': ['<a href=\"https://maps.
                    google.com/maps/contrib/987654321\">Chengdu Education Center</a
                    >'], 'photo_reference': 'XYZ987', 'width': 600}], 'place_id': '
                    ChIJ9876543210', 'plus_code': {'compound_code': '9X7W+Q9 Chengdu
                    , Sichuan, China', 'global_code': '8Q7X9X7W+Q9'}, 'rating': 4.7,
                     'reference': 'XYZ987', 'types': ['education', '
                    point_of_interest', 'establishment'], 'user_ratings_total':
                    456}]})"
            }
        }
    }
]
```

Listing 7: Example of planning process.(Health-PAPI)

```
[
    {
        "MessageNumber": "1241",
        "Query": "\"First, I want to check the rating of my favorite movie
            Interstellar, then see if there is an official trailer available, get
            the genre and user rating of the movie from Netflix, and finally use
            IMDBAPI to check the availability of the movie in different countries
            .\""
    },
    {
        "GlobalThought": "Develop a comprehensive plan to gather detailed
            information about the movie 'Interstellar', including its ratings,
            trailer availability, genre, and Netflix availability across different
            countries.",
        "OrderSteps": {
            "TotalSteps": 4,
            "StepDetail": {
                "StepNumber": 1,
                "Description": "Check the rating of the movie 'Interstellar'",
                "Action": "IMDBAPI({'query': 'Interstellar'})",
                "Results": "IMDBAPI({'d': [{'id': 'tt0816692', 'l': 'Interstellar',
                    'q': 'feature', 'rank': '28', 's': 'Matthew McConaughey, Anne
                    Hathaway', 'v': [{'id': 'vi1586278169', 'l': 'Interstellar
                    Official Trailer #1', 's': '2:51'}, {'id': 'vi1586278170', 'l':
```

```
                    'Interstellar Official Trailer #2', 's': '2:30'}], 'vt': '2014',
                     'y': '2014', 'yr': '2014-2014'}]})"
            }
        }
    },
    {
        "GlobalThought": "Develop a comprehensive plan to gather detailed
            information about the movie 'Interstellar', including its ratings,
            trailer availability, genre, and Netflix availability across different
            countries.",
        "OrderSteps": {
            "TotalSteps": 4,
            "StepDetail": {
                "StepNumber": 2,
                "Description": "Check if there is an official trailer available for
                    'Interstellar'",
                "Action": "YouTubeDataAPI({'q': 'Interstellar Official Trailer', 'hl
                    ': 'en', 'gl': 'US'})",
                "Results": "YouTubeDataAPI({'query': 'Interstellar Official Trailer
                    ', 'results': ['Interstellar Official Trailer #1', 'Interstellar
                     Official Trailer #2', 'Interstellar - Official Trailer [HD]', '
                    Interstellar - Trailer 2']})"
            }
        }
    },
    {
        "GlobalThought": "Develop a comprehensive plan to gather detailed
            information about the movie 'Interstellar', including its ratings,
            trailer availability, genre, and Netflix availability across different
            countries.",
        "OrderSteps": {
            "TotalSteps": 4,
            "StepDetail": {
                "StepNumber": 3,
                "Description": "Get the genre and user rating of 'Interstellar' from
                     Netflix",
                "Action": "uNoGSAPI({'title': 'Interstellar', 'type': 'movie'})",
                "Results": "uNoGSAPI({'results': [{'country': 'United States', '
                    countrycode': 'US', 'expiring': 0, 'id': 123456, 'nl7': 1, '
                    tmovs': 1, 'tseries': 0, 'tvids': 0}]})"
            }
        }
    },
    {
        "GlobalThought": "Develop a comprehensive plan to gather detailed
            information about the movie 'Interstellar', including its ratings,
            trailer availability, genre, and Netflix availability across different
            countries.",
        "OrderSteps": {
            "TotalSteps": 4,
            "StepDetail": {
                "StepNumber": 4,
                "Description": "Check the availability of 'Interstellar' in
                    different countries using IMDBAPI",
                "Action": "IMDBAPI({'query': 'Interstellar', 'country_list': 'US,CN,
                    IN,JP,UK'})",
                "Results": "IMDBAPI({'d': [{'id': 'tt0816692', 'l': 'Interstellar',
                    'q': 'feature', 'rank': '28', 's': 'Matthew McConaughey, Anne
                    Hathaway', 'v': [{'id': 'vi1586278169', 'l': 'Interstellar
                    Official Trailer #1', 's': '2:51'}, {'id': 'vi1586278170', 'l':
                    'Interstellar Official Trailer #2', 's': '2:30'}], 'vt': '2014',
                     'y': '2014', 'yr': '2014-2014', 'countryAvailability': ['US', '
                    CN', 'IN', 'JP', 'UK']}]})"
            }
        }
    }
]
```


Listing 8: Example of planning process.(Entertainment-CAPI)

```
[
    {
        "MessageNumber": "2604",
        "Query": "\"Can I use Coursera to register for a series of related Japanese
            courses, actively participate in discussion forums, complete all
            assignments on time, obtain certificates for each course, use Google
            Calendar to manage my study plans and deadlines, and use WhatsApp to
            collaborate with classmates?\""
    },
    {
        "GlobalThought": "Develop a strategic plan to effectively use Coursera for
            Japanese courses, manage study schedules with Google Calendar, and
            collaborate with classmates via WhatsApp.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 1,
                "Description": "Open Coursera app",
                "Action": {
                    "Tap": "Tap on the Coursera app icon to open it."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "AppOpened": "Coursera app successfully opened.",
                        "HomeScreen": "Navigated to the home screen displaying
                            available courses."
                    }
                }
            }
        }
    },
    {
        "GlobalThought": "Develop a strategic plan to effectively use Coursera for
            Japanese courses, manage study schedules with Google Calendar, and
            collaborate with classmates via WhatsApp.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 2,
                "Description": "Search for Japanese courses",
                "Action": {
                    "Tap": "Tap on the search bar and type 'Japanese courses', then
                        tap the search icon."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "SearchInitiated": "Search bar activated and query entered
                            .",
                        "SearchResults": "Displayed a list of Japanese courses
                            matching the search criteria."
                    }
                }
            }
        }
    },
    {
        "GlobalThought": "Develop a strategic plan to effectively use Coursera for
            Japanese courses, manage study schedules with Google Calendar, and
            collaborate with classmates via WhatsApp.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 3,
                "Description": "Select a Japanese course series",
                "Action": {
                    "Tap": "Tap on a series of related Japanese courses from the
                        search results to view details."
```

```
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "CourseDetailsDisplayed": "Details of selected Japanese
                            courses are displayed.",
                        "CourseInformation": {
                            "Course1": {
                                "Title": "Japanese for Beginners",
                                "Instructor": "Dr. Yuki Tanaka",
                                "Duration": "6 weeks",
                                "Rating": "4.7"
                            },
                            "Course2": {
                                "Title": "Intermediate Japanese",
                                "Instructor": "Prof. Hiroshi Nakamura",
                                "Duration": "8 weeks",
                                "Rating": "4.5"
                            }
                        }
                    }
                }
            }
        }
    },
    {
        "GlobalThought": "Develop a strategic plan to effectively use Coursera for
            Japanese courses, manage study schedules with Google Calendar, and
            collaborate with classmates via WhatsApp.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 4,
                "Description": "Enroll in the selected courses",
                "Action": {
                    "Tap": "Tap the 'Enroll' button for each course in the series to
                         register."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "EnrollmentStatus": {
                            "Course1": {
                                "Title": "Japanese for Beginners",
                                "EnrollmentStatus": "Enrolled",
                                "Confirmation": "You are now enrolled in 'Japanese
                                    for Beginners'."
                            },
                            "Course2": {
                                "Title": "Intermediate Japanese",
                                "EnrollmentStatus": "Enrolled",
                                "Confirmation": "You are now enrolled in '
                                    Intermediate Japanese'."
                            }
                        }
                    }
                }
            }
        }
    },
    {
        "GlobalThought": "Develop a strategic plan to effectively use Coursera for
            Japanese courses, manage study schedules with Google Calendar, and
            collaborate with classmates via WhatsApp.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 5,
                "Description": "Open Google Calendar to manage study plans",
                "Action": {
```

```
                    "Tap": "Tap on the Google Calendar app icon to open it."
                },
                "Results": {
                    "Status": "Success",
                    "Response": {
                        "AppOpened": "Google Calendar app successfully opened.",
                        "CalendarView": "Displayed the current month's calendar with
                             scheduled events."
                    }
                }
            }
        }
    }
]
```

Listing 9: Example of planning process.(Edu-CAPP)

```
[
    {
        "MessageNumber": "3602",
        "Query": "What's the weather like in Sanya next Friday What are the famous
            local snacks in Sanya Reserve a restaurant at 930 AM in Tianyahaijiao
            that provides my favorite food Reserve my favorite hotel at 345 PM in
            Yalong Bay Beach Save this travel and activity plan to the memo"
    },
    {
        "GlobalThought": "Plan a travel and activity itinerary for Sanya, including
            checking the weather, finding local snacks, booking a restaurant and
            hotel, and saving the plan to the memo.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 1,
                "Description": "Query the weather in Sanya next Friday",
                "Action": "queryWeather({'date': '2024-12-13', 'city': 'Sanya'})",
                "Results": "queryWeather({'weather': 'sunny', 'date': '2024-12-13',
                    'city': 'Sanya', 'minCelsius': 22.5, 'maxCelsius': 28.0})"
            }
        }
    },
    {
        "GlobalThought": "Plan a travel and activity itinerary for Sanya, including
            checking the weather, finding local snacks, booking a restaurant and
            hotel, and saving the plan to the memo.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 2,
                "Description": "Find famous local snacks in Sanya",
                "Action": "getCitySpecialties({'city': 'Sanya', 'foodType': 'snacks
                    '})",
                "Results": "getCitySpecialties({'food': 'Coconut Candy', 'foodType':
                     'snacks', 'averagePriceOfFood': '15 CNY', 'calorie': 350.0, '
                    healthLevel': 4})"
            }
        }
    },
    {
        "GlobalThought": "Plan a travel and activity itinerary for Sanya, including
            checking the weather, finding local snacks, booking a restaurant and
            hotel, and saving the plan to the memo.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 3,
                "Description": "Book a restaurant at Tianyahaijiao that provides
                    favorite food",
                "Action": "bookRestaurant({'reservationDate': '2024-12-13', '
                    reservationTime': '09:30', 'restaurantLocation': 'Tianyahaijiao,
                     Sanya', 'city': 'Sanya', 'userContactNumber': '13758294715', '
```

```
                peopleNumber': 2, 'restaurantName': 'Shu Jiu Xiang Hotpot', '
                providingFood': 'Soup Dumplings', 'specialRequests': 'Provide a
                portable charger'})",
            "Results": "bookRestaurant({'reservationDate': '2024-12-13', '
                reservationTime': '09:30', 'restaurantName': 'Shu Jiu Xiang
                Hotpot', 'restaurantType': 'Hotpot', 'location': 'Tianyahaijiao,
                 Sanya', 'userContactNumber': '13758294715', 'peopleNumber': 2,
                'providingFood': 'Soup Dumplings', 'specialRequests': 'Portable
                charger provided'})"
        }
    }
},
{
    "GlobalThought": "Plan a travel and activity itinerary for Sanya, including
        checking the weather, finding local snacks, booking a restaurant and
        hotel, and saving the plan to the memo.",
    "OrderSteps": {
        "TotalSteps": 5,
        "StepDetail": {
            "StepNumber": 4,
            "Description": "Reserve the favorite hotel at Yalong Bay Beach",
            "Action": "bookHotel({'hotelName': 'Atour Hotel', 'checkInDate':
                '2024-12-13', 'checkInTime': '15:45', 'checkOutDate':
                '2024-12-14', 'roomType': 'Business King Room', 'numberOfRooms':
                 1, 'peopleNumber': 2, 'userContactNumber': '13758294715', '
                specialRequests': 'Room near the swimming pool'})",
            "Results": "bookHotel({'hotelName': 'Atour Hotel', 'location': '
                Sanya', 'checkInDate': '2024-12-13', 'checkInTime': '15:45', '
                checkOutDate': '2024-12-14', 'checkOutTime': '12:00', '
                userContactNumber': '13758294715', 'roomType': 'Business King
                Room', 'numberOfRooms': 1, 'peopleNumber': 2, 'totalPrice': '680
                 CNY', 'specialRequests': 'Room near the swimming pool confirmed
                '})"
        }
    }
},
{
    "GlobalThought": "Plan a travel and activity itinerary for Sanya, including
        checking the weather, finding local snacks, booking a restaurant and
        hotel, and saving the plan to the memo.",
    "OrderSteps": {
        "TotalSteps": 5,
        "StepDetail": {
            "StepNumber": 5,
            "Description": "Save this travel and activity plan to the memo",
            "Action": "saveToMemo({'startDate': '2024-12-13', 'endDate':
                '2024-12-14', 'dateAndWeather': '2024-12-13 : sunny, minCelsius:
                 22.5, maxCelsius: 28.0', 'cityAndSpecialties': 'Sanya : Coconut
                 Candy', 'dateAndAction': '2024-12-13 : Travel to Sanya', '
                restaurantName': 'Shu Jiu Xiang Hotpot', 'hotelName': 'Atour
                Hotel', 'hotelPrice': '680 CNY', 'totalPrice': '680 CNY'})",
            "Results": "saveToMemo({'startDate': '2024-12-13', 'endDate':
                '2024-12-14', 'cities': 'Sanya', 'specialtiesByCity': 'Coconut
                Candy', 'hotelName': 'Atour Hotel', 'hotelPrice': 680.0, '
                restaurantName': 'Shu Jiu Xiang Hotpot', 'totalPrice': 680.0})"
        }
    }
}
]
```

Listing 10: Example of planning process.(Travel-DAG)

```
[
{
    "MessageNumber": "4213",
    "Query": "Please help me search for a pair of yoga pants on my usual
        platform, with a budget of 300 yuan, free shipping and same-day delivery
        . Add this item to the shopping cart. Then proceed to pay for this pair
        of yoga pants. Use the same conditions to search for the item on Taobao.
```

```
            Please determine which one has the highest cost-performance ratio among
            these two items."
    },
    {
        "GlobalThought": "Search for yoga pants within budget on Pinduoduo and
            Taobao, add to cart, pay, and compare cost-effectiveness.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 1,
                "Description": "Search for yoga pants on Pinduoduo",
                "Action": "searchGoodsOnline({'goodName': 'yoga pants', 'platform':
                    'Pinduoduo', 'maxPrice': 300, 'sameDayDelivery': true, '
                    freeShipping': true})",
                "Results": "searchGoodsOnline({'goodId': 'PD123456', 'goodName': '
                    yoga pants', 'goodCategory': 'Apparel', 'singlePrice': 250, '
                    platform': 'Pinduoduo', 'sameDayDelivery': true, 'brand': '
                    SportswearBrand', 'freeShipping': true, 'shippingCost': 0})"
            }
        }
    },
    {
        "GlobalThought": "Search for yoga pants within budget on Pinduoduo and
            Taobao, add to cart, pay, and compare cost-effectiveness.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 2,
                "Description": "Add the selected yoga pants from Pinduoduo to the
                    shopping cart",
                "Action": "addToCart({'goodId': 'PD123456', 'quantity': 1, 'goodName
                    ': 'yoga pants', 'goodSinglePrice': 250, 'userContactNumber':
                    '13812345678'})",
                "Results": "addToCart({'result': [{'goodId': 'PD123456', 'quantity':
                     1, 'goodName': 'yoga pants'}]})"
            }
        }
    },
    {
        "GlobalThought": "Search for yoga pants within budget on Pinduoduo and
            Taobao, add to cart, pay, and compare cost-effectiveness.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 3,
                "Description": "Proceed to pay for the yoga pants from Pinduoduo",
                "Action": "buyGood({'goodId': 'PD123456', 'quantity': 1, '
                    pricePerUnit': 250, 'shippingCost': 0, 'totalPrice': 250, '
                    contactNumber': '13812345678', 'address': 'Room 1202, Building
                    5, No. 88 Haier Road, Laoshan District, Qingdao', 'paymentMethod
                    ': 'Alipay'})",
                "Results": "buyGood({'status': 'success', 'goodId': 'PD123456', '
                    quantity': 1, 'pricePerUnit': 250, 'promoCode': '', '
                    shippingCost': 0, 'finalPrice': 250, 'platform': 'Pinduoduo', '
                    paymentMethod': 'Alipay', 'contactNumber': '13812345678', '
                    address': 'Room 1202, Building 5, No. 88 Haier Road, Laoshan
                    District, Qingdao'})"
            }
        }
    },
    {
        "GlobalThought": "Search for yoga pants within budget on Pinduoduo and
            Taobao, add to cart, pay, and compare cost-effectiveness.",
        "OrderSteps": {
            "TotalSteps": 5,
            "StepDetail": {
                "StepNumber": 4,
                "Description": "Search for yoga pants on Taobao",
                "Action": "searchGoodsOnline({'goodName': 'yoga pants', 'platform':
                    'Taobao', 'maxPrice': 300, 'sameDayDelivery': true, '
```

```
                freeShipping': true})",
            "Results": "searchGoodsOnline({'goodId': 'TB789101', 'goodName': '
                yoga pants', 'goodCategory': 'Apparel', 'singlePrice': 280, '
                platform': 'Taobao', 'sameDayDelivery': true, 'brand': '
                FitnessWearBrand', 'freeShipping': true, 'shippingCost': 0})"
        }
    }
},
{
    "GlobalThought": "Search for yoga pants within budget on Pinduoduo and
        Taobao, add to cart, pay, and compare cost-effectiveness.",
    "OrderSteps": {
        "TotalSteps": 5,
        "StepDetail": {
            "StepNumber": 5,
            "Description": "Compare the cost-effectiveness of yoga pants from
                Pinduoduo and Taobao",
            "Action": "compareCostEffectiveness({'product1Name': 'yoga pants', '
                product2Name': 'yoga pants', 'product1Price': 250, '
                product2Price': 280, 'product1Rating': 4.7, 'product2Rating':
                4.6})",
            "Results": "compareCostEffectiveness({'betterGoodId': 'PD123456', '
                betterGoodName': 'yoga pants', 'betterGoodPrice': 250, '
                betterGoodRating': 4.7})"
        }
    }
}
]
```

Listing 11: Example of planning process.(Shopping-DAG)

Scenario	Task
Travel	I want to know the weather in Sanya next Friday. Please check the train tickets to Sanya on that day. Also, check the flight tickets to Sanya on that day. Reserve my favorite restaurant and my favorite seat at Tianya-haijiao Square at 9:30 AM. Reserve my favorite hotel room with my usual request at Yalong Bay Beach at 3:45 PM.
Entertainment	I want to know what the weather will be like in Zhuhai City in 7 days. On that day, what are the scheduled light show activities in Zhuhai City? If there are tickets available, please book one for me. I also want to know what the specialties of Zhuhai City are? Finally, 2 hours after the event ends, help me reserve a restaurant at the Sun and Moon Shell Theater that serves these specialties.
Shopping	Please help me search for a pair of white high heels on my favorite platform, with a budget of 800 yuan, and it should support returns. Add this item to the shopping cart. Then pay for this pair of white high heels. Use the same conditions to search for this item on Suning.com. Please find out which one has the highest cost-performance ratio among these two items.
Edu	Please help me find the basic entry-level course of the course I am studying on my favorite education platform, Wangyi Cloud Classroom. If it exists, please register me for it. Then register me for the advanced comprehensive level course of this course. Also, register me for the basic entry-level course of this course on Tencent Class. And register me for the advanced comprehensive level course of this course on Tencent Class. Finally, what courses have been registered in total.

Scenario	Task
Health	First, use the ExerciseDB tool to calculate my body mass index BMI. Then, based on the BMI results, use the NutritionCalculator tool to provide personalized diet and exercise recommendations. Next, use the ExerciseDB tool to determine my daily nutritional needs. Finally, use the ExerciseDB tool to provide targeted exercise advice for my health.
Develop	How to set up continuous integration with Travis CI, run unit tests and integration tests, generate detailed test reports, and receive notifications via Slack when the build succeeds or fails? At the same time, how to use GitHub for code hosting and version control, and use JIRA for project management?
Meeting	Use Zoom to schedule an online meeting for quarterly financial review and invite 30 participants. Set the meeting duration to 1.5 hours, enable the waiting room feature, allow participants to record the meeting, set a password, limit each participant's speaking time, and enable automatic mute. Use Microsoft Teams to send reminder messages, use Trello to manage meeting tasks, and automatically generate and share meeting minutes with all participants via Slack using Google Docs after the meeting.
Calendar	Book a training room in Google Calendar for an employee workshop, send meeting invitations via Microsoft Teams, synchronize the meeting time to Outlook Calendar, send reminders via Slack, and test the connection using Zoom one hour before the meeting starts.
Game	How to purchase my favorite game StarCraft on Nintendo eShop, ensuring that the account balance is sufficient or using my most commonly used payment method? If the balance is insufficient, how to recharge? If the recharge fails, what other payment methods can be used?
Diet	Can I use UberEats to find nearby restaurants based on rating 4.8, reviews, and cuisine category spicy hot pot set, add my favorite dish roast duck to my wishlist, verify the restaurant location with Google Maps, complete payment via PayPal, set a pickup time reminder at 6:00 PM with Google Calendar, and update payment methods and preferences in my user profile?

Table 10: Task cases in different scenarios.