# AURA: A Diagnostic Framework for Tracking User Satisfaction of Interactive Planning Agents

**Takyoung Kim*** **Janvijay Singh*** **Shuhaib Mehri***
**Emre Can Acikgoz** **Sagnik Mukherjee** **Nimet Beyza Bozdag** **Sumuk Shashidhar**
**Gokhan Tur** **Dilek Hakkani-Tür**
University of Illinois Urbana-Champaign
{tk30, jvsingh2, mehri2, gokhan, dilek}@illinois.edu

## Abstract

The growing capabilities of large language models (LLMs) in instruction-following and context-understanding lead to the era of agents with numerous applications. Among these, task planning agents have become especially prominent in realistic scenarios involving complex internal pipelines, such as context understanding, tool management, and response generation. However, existing benchmarks predominantly evaluate agent performance based on task completion as a proxy for overall effectiveness. We hypothesize that merely improving task completion is misaligned with maximizing user satisfaction, as users interact with the entire agentic process and not only the end result. To address this gap, we propose **AURA**, an Agent-User inteRaction Assessment framework that conceptualizes the behavioral stages of interactive task planning agents. AURA offers a comprehensive assessment of agent through a set of atomic LLM evaluation criteria, allowing researchers and practitioners to diagnose specific strengths and weaknesses within the agent's decision-making pipeline. Our analyses show that agents excel in different behavioral stages, with user satisfaction shaped by both outcomes and intermediate behaviors. We also highlight future directions, including systems that leverage multiple agents and the limitations of user simulators in task planning.

## 1 Introduction

Large language models (LLMs) are increasingly deployed in real-world applications, primarily due to their ability to understand complex goals and devise structured sequences of action: a process known as "planning" [27]. To assess and refine planning skills, researchers have introduced diverse benchmarks in web-based [35, 34], mobile [8], embodied [7], and automated testbeds [19].

One of the most impactful planning applications is *task planning*, where agents generate and execute domain-specific plans (*e.g.*, itineraries) to help users achieve goals while adhering to contextual constraints [4, 20, 33]. While agents must accurately interpret user requests, leverage predefined tools, and engage in personalized dialogues, many benchmarks focus solely on final task completion, overlooking the agents' intermediate behaviors and planning steps. Given that **overall user satisfaction is shaped by multiple factors** (*e.g.*, efficiency and effectiveness) **throughout prolonged interactions** [3], a narrow focus on outcome-based metrics can misrepresent an agent's true effectiveness, especially in complex task-oriented applications [6, 31, 1].

In addition, existing benchmarks across different domains often introduce *bespoke* evaluation frameworks and metrics, making it possible for the same agent to be evaluated using entirely different

[*] Equal Contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

criteria. This fragmentation introduces inconsistencies in how agent behavior is interpreted and makes it difficult to draw generalizable conclusions about agent capabilities. Without a domain-agnostic evaluation framework, it becomes challenging to meaningfully compare systems, track progress over time, and identify fundamental limitations in agent design.

To address these limitations, we propose **AURA**, an Agent-User inteRaction Assessment framework. To ensure generalizability across sequential agentic scenarios, AURA is designed based on the partially observed Markov Decision Process (POMDP). This formulation reflects the reality that key aspects of interaction, such as user intention and satisfaction, are often hidden or only partially observable. Moreover, it aligns with the observation that contemporary agent behaviors typically unfold in a sequential, decision-making context consistent with the POMDP paradigm [25, 11, 33, 36]. Lastly, AURA defines domain-agnostic LLM evaluation criteria that can be easily instantiated through a set of atomic and easily measurable metrics. As illustrated in Figure 1, AURA provides the following key advantages:

1. **Generalizable Evaluation**: AURA establishes a domain-agnostic protocol for interactive task planning, enabling a consistent comparison across diverse benchmarks under a shared set of principles, which has been discussed for decades [26] yet remains an open issue.
2. **Multi-Axis Diagnosis**: By aligning with the POMDP framework, AURA supports evaluation across multiple behavioral stages of the interaction pipeline, encompassing both intermediate decisions and final outcomes. This allows for a fine-grained diagnostic analysis of how different agents behave and impact performance.
3. **Cross-Benchmark Comparisons**: AURA enables systematic comparisons across heterogeneous tasks and environments, thereby uncovering planning strategies and revealing performance trade-offs that may not be observable within the scope of individual benchmarks.

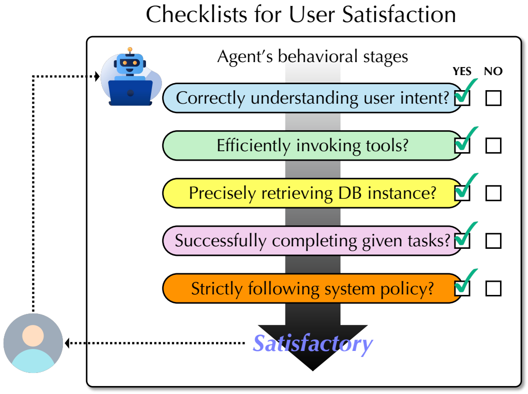

Checklists for User Satisfaction
Agent's behavioral stages
YES
NO
Correctly understanding user intent?
Efficiently invoking tools?
Precisely retrieving DB instance?
Successfully completing given tasks?
Strictly following system policy?
Satisfactory

Figure 1: AURA provides unified, atomic, and domain-agnostic criteria for assessing user satisfaction of interactive planning agents, extending beyond conventional evaluation protocols that focus solely on task completion.

Through extensive experiments, we demonstrate that different models exhibit distinct strengths and weaknesses within decision-making pipelines. Furthermore, human studies indicate that stage-specific evaluation using AURA correlates more strongly with improvements in user satisfaction than with the final task completion metric alone. Lastly, analyses on combining agents during deployment and the reliability of user simulators highlight promising research directions for future scholars and practitioners.

## 2 Background: Bespoke Evaluation of Planning Agents

Recent benchmarks for evaluating agentic tasks adopt rigid, domain-specific metrics to reflect the unique demands of each task setting, as summarized in Appendix A. This specialization is often necessary: interactive planning tasks vary widely in structure, goals, and interaction modalities, requiring tailored criteria to capture meaningful performance signals.

For instance, AgentBench [16] incorporates domain-specific metrics such as Success Rate, Win Rate, F1 Score, and Exact Match to assess agents comprehensively. FlowBench [32] further enhances the evaluation landscape by adding metrics focused explicitly on tool usage: Tool Invocation, measured by precision, recall, and F1 score for identifying the correct tool configurations; Success Rate, denoting the proportion of entirely successful sessions; and Task Progress, capturing the percentage of goals completed within a session. Meanwhile, $\tau$-Bench [36] utilizes more direct measures of success with Pass@k, a boolean indicator reflecting success across k attempts, and Pass^k, which requires successful outcomes for all k trials.

These metrics are often rigid by design, optimized for narrowly defined outcomes within their respective domains. However, this tight coupling limits their generalizability: it is difficult to apply these metrics across tasks or to capture broader notions of agent capability, especially in open-ended or compositional settings. As noted in prior work [21], overemphasis on final task success obscures nuanced failures or partial progress, particularly problematic in subjective or exploratory tasks like information aggregation [9, 38]. **In contrast, AURA seeks to define a more general and domain-agnostic evaluation paradigm that is not tied to narrow success criteria.**

## 3 Method

As a prerequisite step, we set minimum requirements for task planning benchmarks to effectively show realistic scenarios. Specifically, task planning benchmarks should define specific **tools** (*e.g.*, APIs) and construct domain-specific **databases**. These assumptions are adopted because their dedicated environmental constraints present challenges in achieving a unified evaluation.

### 3.1 Metric Design Criteria

Interactive planning tasks are generally modeled using the POMDP paradigm, where agents devise plans based on partial observational data [25, 11, 33, 36]. Building on the framework of [29], we present an abstract and discrete pipeline consisting of $\mathcal{S}$ (a set of agent states), $\mathcal{A}$ (a set of actions), $\mathcal{O}$ (a set of observations) and $\mathcal{R}$ (a set of rewards), along with an our additional policy definition $\mathcal{P}$ (a set of global policies). While task completion is a common performance goal, we hypothesize that it is insufficient: user satisfaction depends on *how* an agent arrives at outcomes [3, 24].

To that end, AURA decomposes evaluation into atomic, phase-specific criteria reflecting the multi-stage behavior of LLM agents. These include state consistency, tool efficiency, observation alignment, policy adherence, and task completion, each tied to a specific step in the decision pipeline. Drawing from prior work on decomposed evaluation [17, 22, 13], we design intuitive LLM-based evaluations for each metric for more consistent and precise diagnostics.

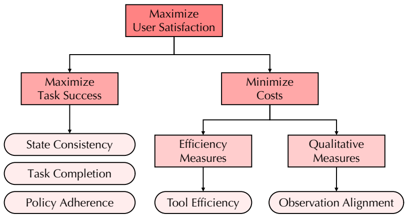


Figure 2: Decision theory-based taxonomy of evaluation metrics in AURA. As discussed in [26], it should be cautious to generalize the original taxonomy to different agents and tasks. Following this, we provide five distinct interpretations of each element, which will be described in Section 3.2.

Furthermore, as illustrated in Figure 2, we align AURA with the decision theoretic principles [26], where disparate metrics collectively estimate user satisfaction. POMDPs provide a natural foundation here, as agents select actions under uncertainty to maximize expected utility, capturing both goal achievement and procedural quality. This integration enables AURA to measure user satisfaction holistically while remaining grounded in theoretical rigor.

Throughout this work, we formulate the basic components of multi-turn, multi-step agents as follows: $T$ denotes the total number of turns (*i.e.*, alternating utterances between the user and the agent) in an interactive session, $M_t$ represents the number of steps in the agent's internal reasoning process within a single utterance at turn $t$, and $\mathcal{C}_t = \{u_\tau, a_\tau\}_{\tau=1}^{t}$ signifies the context of user inputs and agent responses observed up to turn $t$.

### 3.2 Components of AURA

#### 3.2.1 State Consistency ($\mathcal{S}$)

The intermediate decision-making step of the agent is pivotal in summarizing user requests and determining optimal subsequent actions [23, 28]. Given that both pipelined and end-to-end multi-turn interactions are inherently susceptible to error accumulation [15, 14], it is crucial to validate whether intermediate outcomes consistently mediate between user inputs and agent outputs. In this context, the *state consistency* metric measures whether **an agent correctly aligns user requests with its $k$-th intermediate steps ($z_t^k$).** Notably, these intermediate steps can be represented in either a "structured format" (*e.g.*, dialogue states) or "natural language" (*e.g.*, Chain-of-Thought). By accommodating both representations, the state consistency can be formulated as follows:

$$\text{AURA}_{\mathcal{S}} = \frac{\displaystyle\sum_{t=1}^{T}\sum_{k=1}^{M_t} \text{IsConsistent}\big(z_t^k,\ \mathcal{C}_t \backslash \{a_t\}\big)}{\displaystyle\sum_{t=1}^{T} M_t}$$

Here, the `IsConsistent` function serves as a boolean indicator evaluated by LLMs that compares user requests (mostly natural language) with internal states (structured or natural language)[2].

### 3.2.2 Tool Efficiency ($\mathcal{A}$)

The management of external tools or functions, typically in the form of APIs, incurs operational costs and affects the task performance, underscoring the importance of evaluating their effective utilization. In particular, the occurrence of failed API calls prior to collecting complete information from users can result in unnecessary expenditures of resources (*i.e.*, time and money) and negative user experience. To address this, the *tool efficiency* metric first considers **the total number of tool calls ($N_T$)**. For a given task, agent scenarios that require fewer API calls are considered indicative of more efficient API management.

Additionally, due to the stochastic nature of natural language prompting, generating well-structured API calls poses a significant challenge for certain agents. Even when an agent invokes an API at an appropriate time, iterative attempts due to incomplete generation will eventually lead to additional, avoidable costs. Consequently, this metric also calculates **the number of failed tool generation attempts ($N_F$)**. Formally, we propose the following tool efficiency measure:

$$\text{AURA}_{\mathcal{A}} = \frac{N_T - N_F}{N_T + N_F}$$

where the numerator rewards successful calls and the denominator penalizes excessive or failed calls. A higher value of $\text{AURA}_{\mathcal{A}}$ indicates more efficient tool usage: maximizing successful calls while minimizing total and failed calls.

### 3.2.3 Observation Alignment ($\mathcal{O}$)

The *observation alignment* metric evaluates **whether observations appearing within the context align with what the user requires.** Specifically, it is calculated with a boolean criteria for each observation (*i.e.*, retrieved database entity), preceded by capturing the number of observations ($|O|$) explicitly appear within the agent responses. Considering the fact that both user utterance and agent response are represented in natural languages, it can be measured via a set of atomic LLM evaluations:

$$\text{AURA}_{\mathcal{O}} = \frac{1}{|T_{\text{obs}}|} \sum_{t \in T_{\text{obs}}} \left( \frac{1}{|O^{(t)}|} \sum_{o \in O^{(t)}} \text{IsAligned}(o, \mathcal{C}_t \backslash \{a_t\}) \right) \text{, where } T_{obs} = \{t \in T : |O^{(t)}| > 0\}$$

Intuitively, $T_{obs}$ denotes a set of turns where observations are present within agent responses ($|O| > 0$), and observations of a specific turn $t \in T_{obs}$ are extracted from LLM (denoted $O^{(t)}$). In addition, `IsAligned` serves as a boolean indicator assessing whether each observation aligns with conversational context. By quantifying and improving observation alignment, we not only promote the clarity of agent outputs but also minimize the system's overall costs by decreasing the possibility of repetitively calling tools.

[2] Practically, as observed by [13], evaluating a group of targeted samples collectively, using simple and atomic criteria such as boolean assessments, does not adversely affect performance compared to stepwise evaluation. Therefore, we group $z_t^{1:k}$ in the implementation.

Table 1: Benchmark statistics. We use a validation set for TravelPlanner that demonstrates a similar performance pattern with the test set.

| | # of Scenarios | # of Tools | # of Database |
|---|---|---|---|
| TravelPlanner [33] | 180 | 7 | 3,865,195 total, 3,827,361 max for a tool |
| $\tau$-Bench-Airline [36] | 50 | 13 | 500 users, 300 flights, 2,000 reservations |
| $\tau$-Bench-Retail [36] | 115 | 15 | 500 users, 50 products, 1,000 orders |

### 3.2.4 Policy Alignment ($\mathcal{P}$)

Interactive agent benchmarks assume certain policies (*i.e.*, a behavioral rule appearing in agent responses, not only emerging in their internal states) that are globally reflected across the interactive sessions [32], mostly in a form of system prompt. The *Policy alignment* is a session-level metric measuring whether a predefined set of policies ($\mathcal{P}$) are consistently followed throughout interactive sessions.

$$\text{AURA}_{\mathcal{P}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \text{IsAdherent}(p; \mathcal{C}_T)$$

Similar to the state consistency (Section 3.2.1) and observation alignment (Section 3.2.3) metrics, `IsAdherent` serves as an LLM-evaluated boolean indicator deciding consistency between each policy and interaction context.

### 3.2.5 Task Completion ($\mathcal{R}$)

As in many agent studies, the primary objective of a task agent is to effectively accomplish goal-oriented tasks. The *task completion* metric adheres to the task-specific performance criteria established by each benchmark, since benchmark scenarios typically define their own evaluation frameworks to determine what constitutes a "completed task." Specifically, most metrics in diverse benchmarks demonstrated in Table 5 can be regarded as task completion metrics since they mainly focus on whether agents successfully accomplish the given task. Again, it is important to note that **AURA incorporates existing task completion metrics** to complement the holistic evaluation, rather than excluding them.

### 3.2.6 Agent Interaction Pattern

While prior works have typically reported on dataset-level statistics (*e.g.*, basic information in Table 1), it has rarely addressed how agents actually interact within their environments. We additionally report the following information, providing an overview of the interactive tendency of employed agents in their specific scenarios.

**The Number of Turns:** Different from offline dialogue datasets where a deterministic number of turns is provided, recently employed online evaluation scenarios typically omit details about the agent interactivity. However, the number of interactive turns serves as a critical indicator of the conversational or interactive dynamics. In general, planning scenarios with more turns tend to require further management of user constraints, suggesting the need for sustained engagement and state tracking.

**The Number of Steps:** Steps refer to internal actions or reasoning hops, such as goal decomposition or tool use. This feature reflects the agent's ability to handle complex tasks that require breaking down overarching goals into smaller, manageable actions. Fewer steps may suggest shallow or generic reasoning, while more steps indicate finer control. Correlating the count of steps with performance and the length of interaction can reveal whether agents are making meaningful progress or getting stuck in unproductive loops.

Table 2: AURA evaluation result. The relative performance ranking among agents for each metric is differentiated with colors (*i.e.*, darker color indicates more competitive performance; different colors are applied for proprietary and open-source models). Best average performance is indicated in **bold**, and second-best is underlined.

| | TravelPlanner | | | | | | $\tau$-Bench-Airline | | | | | | $\tau$-Bench-Retail | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Agent | $\mathcal{S}$ | $\mathcal{A}$ | $\mathcal{O}$ | $\mathcal{P}$ | $\mathcal{R}$ | AVG. | $\mathcal{S}$ | $\mathcal{A}$ | $\mathcal{O}$ | $\mathcal{P}$ | $\mathcal{R}$ | AVG. | $\mathcal{S}$ | $\mathcal{A}$ | $\mathcal{O}$ | $\mathcal{P}$ | $\mathcal{R}$ | AVG. |
| ***Proprietary Models*** | | | | | | | | | | | | | | | | | | |
| gpt-4o | .98 | .79 | .01 | .48 | .01 | .44 | .58 | .96 | .73 | .83 | .42 | .70 | .53 | .93 | .79 | .89 | .54 | .74 |
| gpt-4o-mini | .94 | .95 | .00 | .50 | .00 | .48 | .49 | .92 | .71 | .84 | .26 | .64 | .48 | .89 | .72 | .89 | .46 | .69 |
| gpt-3.5-turbo | .95 | .93 | .00 | .50 | .00 | .48 | .56 | .79 | .65 | .73 | .08 | .56 | .44 | .69 | .57 | .78 | .21 | .54 |
| gemini-1.5-fsh. | .99 | .71 | .00 | .31 | .00 | .40 | .55 | .84 | .66 | .87 | .32 | .65 | .60 | .85 | .72 | .86 | .15 | .64 |
| sonnet-3.5 | .99 | .99 | .01 | .56 | .01 | **.51** | .80 | .97 | .84 | .87 | .46 | **.79** | .73 | .94 | .85 | .91 | .62 | **.81** |
| ***Open-Source Models*** | | | | | | | | | | | | | | | | | | |
| mistral-large | .87 | .77 | .00 | .36 | .00 | .40 | .31 | .96 | .50 | .76 | .26 | .56 | .23 | .91 | .44 | .80 | .34 | .54 |
| mixtral-8x7B | .94 | .77 | .00 | .27 | .00 | .40 | .25 | .85 | .17 | .64 | .28 | .44 | .27 | .50 | .27 | .65 | .05 | .35 |
| llama-3.3-70B | .95 | .96 | .00 | .39 | .00 | .46 | .38 | .90 | .70 | .81 | .30 | **.62** | .32 | .88 | .79 | .87 | .36 | **.64** |
| qwen2.5-72B | .98 | .97 | .00 | .39 | .00 | **.47** | .35 | .90 | .58 | .79 | .20 | .56 | .27 | .75 | .53 | .84 | .38 | .55 |

# 4 Experiments

## 4.1 Benchmarks

To verify our hypothesis on the evaluation following behavioral stages, we employ challenging task planning benchmarks and compare agent performance based solely on task completion with that achieved using AURA. Specifically, we utilize three task domains of two benchmarks: TravelPlanner [33], a single-turn dataset in the itinerary domain, and $\tau$-Bench [36], a multi-turn dataset encompassing airline and retail domains. Table 1 presents the statistics of these benchmarks. These benchmarks are selected due to their rich environmental constraints that conform to the prerequisite assumptions described in Section 3 (*i.e.*, predefined tools and external database). For the evaluation metric for task completion ($\mathcal{R}$), pass rate is used for TravelPlanner, and pass^$k$ is used in $\tau$-Bench. Refer to Table 5 for the description of each metric.

## 4.2 LLMs for Agent, User Simulator, and Task Evaluator

We test diverse task planning agents consisting of five proprietary models (gpt-4o, gpt-4o-mini, gpt-3.5-turbo, gemini-1.5-flash, sonnet-3.5) and four open-weight models (mistral-large[3], mixtral-8x7B, llama-3.3-70B, qwen2.5-72B), with a temperature of 0.0, respectively. Since $\tau$-Bench requires multi-turn interactions between a user and agent, we employ gpt-4o as a user simulator with the same instruction as the original work [36]. For the LLM evaluator leveraged in measuring AURA metrics, we employ llama-3.3-70B [10][4]. All prompts used in our experiments are listed in Appendix B.

## 4.3 Results

**Behavioral Stage Diagnosis:** Table 2 presents the evaluation results following AURA (see Appendix C for agents' erroneous behaviors at each stage). We differentiate the color ranking of proprietary and open-weight models to visually observe patterns. Along with individual metrics in AURA, we report the average score to see the overall behavioral performance.

A notable observation is that **the conventional task completion metric ($\mathcal{R}$) does not necessarily reflect the performance of intermediate phases.** For instance, while qwen2.5-72B achieves the highest performance on task completion (.38) among open-weight models in $\tau$-Bench-Retail,

[3] 123B-sized model: https://huggingface.co/mistralai/Mistral-Large-Instruct-2411

[4] We confirm llama-3.3-70B shows better performance compared to gpt-4o-mini through our preliminary investigation.

Table 3: The average number of turns and steps within a single interactive session (the agent interaction pattern defined in Section 3.2.6). Note that `Avg. Steps` denotes the number of internal agentic processes in each turn not visible in superficial interactions.

| | TravelPlanner | | $\tau$-Bench-Airline | | $\tau$-Bench-Retail | |
|---|---|---|---|---|---|---|
| Agent | Avg. Turns | Avg. Steps | Avg. Turns | Avg. Steps | Avg. Turns | Avg. Steps |
| ***Proprietary Models*** | | | | | | |
| `gpt-4o` | $1.00_{\pm 0.00}$ | $18.71_{\pm 6.91}$ | $7.70_{\pm 3.08}$ | $0.93_{\pm 1.71}$ | $8.26_{\pm 2.54}$ | $0.98_{\pm 1.23}$ |
| `gpt-4o-mini` | $1.00_{\pm 0.00}$ | $15.16_{\pm 6.67}$ | $8.72_{\pm 3.77}$ | $0.81_{\pm 1.79}$ | $8.52_{\pm 3.09}$ | $1.07_{\pm 1.50}$ |
| `gpt-3.5-turbo` | $1.00_{\pm 0.00}$ | $12.87_{\pm 6.44}$ | $7.66_{\pm 3.70}$ | $0.94_{\pm 1.12}$ | $7.97_{\pm 2.50}$ | $1.17_{\pm 1.19}$ |
| `gemini-1.5-fsh.` | $1.00_{\pm 0.00}$ | $16.57_{\pm 6.97}$ | $3.04_{\pm 2.14}$ | $0.47_{\pm 0.70}$ | $6.74_{\pm 4.77}$ | $0.44_{\pm 0.68}$ |
| `sonnet-3.5` | $1.00_{\pm 0.00}$ | $18.36_{\pm 6.84}$ | $6.20_{\pm 2.10}$ | $1.03_{\pm 1.58}$ | $7.71_{\pm 2.06}$ | $1.15_{\pm 1.49}$ |
| ***Open-Weight Models*** | | | | | | |
| `mistral-large` | $1.00_{\pm 0.00}$ | $14.16_{\pm 7.48}$ | $8.46_{\pm 6.33}$ | $0.62_{\pm 1.74}$ | $12.51_{\pm 6.44}$ | $0.68_{\pm 0.96}$ |
| `mixtral-8x7B` | $1.00_{\pm 0.00}$ | $18.44_{\pm 7.84}$ | $18.14_{\pm 9.99}$ | $0.01_{\pm 0.12}$ | $16.11_{\pm 9.69}$ | $0.01_{\pm 0.09}$ |
| `llama-3.3-70B` | $1.00_{\pm 0.00}$ | $18.52_{\pm 6.54}$ | $8.02_{\pm 4.33}$ | $1.22_{\pm 1.63}$ | $6.44_{\pm 4.03}$ | $1.31_{\pm 1.31}$ |
| `qwen2.5-72B` | $1.00_{\pm 0.00}$ | $18.20_{\pm 7.36}$ | $18.74_{\pm 9.21}$ | $0.04_{\pm 0.86}$ | $17.10_{\pm 9.73}$ | $0.01_{\pm 0.09}$ |

another AURA results lag behind other models (*e.g.*, `llama-3.3-70B`). Moreover, although most agents in TravelPlanner show the same zero task completion performance[5], each model has different intermediate capability patterns according to AURA metrics. These observed patterns imply a potential discrepancy between task completion performance values and qualitative user preferences, as user satisfaction is influenced not only by task completion but also by the quality of interactions across all phases. We introduce a human study for verifying this discrepancy in Section 5.1.

**Interaction Pattern Diagnosis:** We summarize each agent's interaction pattern in Table 3. While TravelPlanner is characterized by single-turn interactions with numerous intermediate steps prior to producing a final response, $\tau$-Bench exhibits fewer intermediate steps per turn but involves multiple turns. These patterns reveal certain behavioral tendencies among the agents; for instance, **competitive agents (*e.g.*, `sonnet-3.5`, `llama-3.3-70B`) appear to engage in more extensive internal reasoning within individual turns, yet participate in fewer turns overall in multi-turn benchmarks.** This observation provides insightful suggestions on scenario diversification. For example, when adapting TravelPlanner for multi-turn scenarios, careful consideration must be given to balancing the number of turns and steps within each turn. Given that service providers often target different application scenarios, analyzing interaction patterns across benchmarks can inform the strategic selection of those most aligned with specific deployment contexts.

## 5 Analyses and Discussions

### 5.1 Relationship Between Task Completion, AURA, and User Satisfaction

Our hypothesis regarding the results presented in Table 2 is that factors beyond task completion may have significantly influenced overall user satisfaction. To investigate this through a human study, we recruit 16 graduate-level participants, a sample size congruent to prior human-computer interaction research [12, 5]. This study is conducted using two controlled scenarios, described as follows:

**(1) Same $\mathcal{R}$ & Different AVG.** We select `gpt-4o-mini` and `mistral-large` in $\tau$-Bench-Airline, as these two models exhibit identical task completion performance (.26), yet differ in their AURA average scores (.64 and .56, respectively).

**(2) Same AVG. & Different $\mathcal{R}$** We select `gemini-1.5-fsh.` and `llama-3.3-70B` in $\tau$-Bench-Retail, as these two models exhibit identical AURA average score (.64), yet differ in their task completion performance (.15 and .36, respectively).

[5] This low task completion result is consistent with the original paper's results [33].

Table 4: Selected results for mixing agent experiments in $\tau$-Bench-Airline. Agents indicated with "**interm.**" are utilized in an intermediate understanding component. Full results and additional discussions are presented in Table 7 and Section 5.2.

| | $\mathcal{S}$ | $\mathcal{A}$ | $\mathcal{O}$ | $\mathcal{P}$ | $\mathcal{R}$ | AVG. |
|---|---|---|---|---|---|---|
| `mistral-large` | $.383_{\pm 0.038}$ | $.924_{\pm 0.046}$ | $.581_{\pm 0.022}$ | $.777_{\pm 0.005}$ | $\mathbf{.358}_{\pm 0.029}$ | .605 |
| + `qwen2.5-72B` **interm.** | $.417_{\pm 0.031}$ | $.939_{\pm 0.017}$ | $.545_{\pm 0.029}$ | $.809_{\pm 0.008}$ | $.336_{\pm 0.050}$ | **.609** |
| `qwen2.5-72B` | $.403_{\pm 0.021}$ | $.951_{\pm 0.015}$ | $.637_{\pm 0.031}$ | $.821_{\pm 0.008}$ | $.329_{\pm 0.021}$ | .628 |
| + `llama-3.3-70B` **interm.** | $.447_{\pm 0.020}$ | $.958_{\pm 0.014}$ | $.663_{\pm 0.022}$ | $.831_{\pm 0.008}$ | $.272_{\pm 0.051}$ | **.634** |
| + `mistral-large` **interm.** | $.423_{\pm 0.039}$ | $.933_{\pm 0.032}$ | $.578_{\pm 0.050}$ | $.776_{\pm 0.009}$ | $\mathbf{.367}_{\pm 0.013}$ | .615 |
| `llama-3.3-70B` | $.463_{\pm 0.055}$ | $.952_{\pm 0.013}$ | $.643_{\pm 0.035}$ | $.818_{\pm 0.015}$ | $.263_{\pm 0.056}$ | **.628** |
| + `qwen2.5-72B` **interm.** | $.404_{\pm 0.033}$ | $.953_{\pm 0.007}$ | $.599_{\pm 0.016}$ | $.803_{\pm 0.014}$ | $\mathbf{.361}_{\pm 0.019}$ | .624 |

Participants are presented with two multi-turn conversations generated by different agents interacting with a `gpt-4o` user simulator. They are asked to indicate which conversation they find more satisfactory. If both conversations are perceived as equally (un)satisfactory, participants are given the option to select "no preference." Each participant evaluates a total of six conversation pairs (randomly sampled three pairs of conversations, respectively). Detailed instructions are demonstrated in Section D.1, and we provide qualitative analysis for participants' decisions in Section D.2.

The results of the human evaluation are presented in Figure 3. For each scenario, we conduct binomial tests by excluding "no preference" responses, yielding statistically significant results (p-value $< 0.05$ in both cases). In study (1), we observe that participants tend to **favor conversations with higher average score when task completion is held constant**. Conversely, in study (2), **task completion still remains a strong predictor of user satisfaction when the average score is identical**. These findings highlight the complementary roles of general task completion and AURA metrics in capturing different facets of user satisfaction, as demonstrated in Table 2. Our study thus supports the use of AURA as a meaningful proxy for user satisfaction and motivates future research into evaluation frameworks that go beyond simple task outcomes.

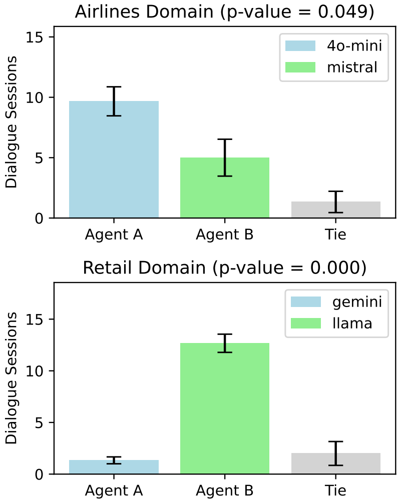


Figure 3: Human study results examining preferred interactions.

## 5.2 Can Mixing Agents Lead to Better Performance?

Building upon the insights from previous experiments, we are further motivated to explore the potential for enhancing performance and user satisfaction by leveraging diverse agents strengths within the agentic pipeline. To facilitate this investigation, we separate the agent's interactive process into two distinct components: (1) intermediate understanding and (2) response generation. This separation enables targeted improvements, wherein state consistency and tool efficiency are addressed in the first stage, whereas observation alignment and policy adherence are the focus of the second. Specifically, we employ one agent to understand user requests and call proper tools, then replace it with the other when starting to generate responses.

For faster investigation of this experimental study, we run FP8-quantized versions of three open-weight models (`llama-3.3-70B`, `qwen2.5-72B`, `mistral-large`), alongside `llama-3.3-70B` as the user simulator, and evaluate on $\tau$-Bench-Airline with five repeated runs per configuration. Selected results are presented in Table 4. **While acknowledging that arbitrary combinations of agents do not consistently yield improved performance, even leading to drop in performance, certain pairings demonstrate notable enhancements in either task completion or the average AURA**

**score**. Notably, combining `qwen2.5-72B` (for intermediate understanding) with `llama-3.3-70B` (for response generation) leads to a substantial improvement in task completion (.263 → .361) without compromising the average AURA score too much (.628 → .624). These findings suggest a promising direction for future research to strategically arrange agents across different stages of the task planning pipeline. For additional details and comprehensive experimental results beyond the scope of this discussion, please refer to Appendix E.

### 5.3 Analysis on the Reliability of User Simulator

To highlight potential limitations of multi-turn evaluation tasks, including those in our own study, we conduct a thorough analysis of user-agent conversations from $\tau$-Bench-Airline, where `gpt-4o` served as the user simulator. We manually examine all user utterances against corresponding user instructions to assess whether the user's behavior aligns with the intended instructions.

Our analysis reveals that in **11 out of 50 conversations (22%), the user simulator demonstrates behaviors not in line with its instructions**. While these deviations do not necessarily lead to failures in task completion, they highlight the importance of considering user simulator performance when evaluating an agent. Although addressing this limitation is left for future work, we identify the following erroneous patterns where a user simulator does not adhere to its instructions: (1) proactivity, (2) instruction contradiction, (3) missing details, and (4) misinterpretation. A comprehensive description of these error categories is provided in Appendix F.

## 6 Conclusion

We propose AURA, a task-agnostic evaluation framework based on a partially observable Markov Decision Process (POMDP), to more accurately capture how user satisfaction emerges from both intermediate behaviors and final outcomes in interactive task planning. Unlike traditional metrics that focus solely on task completion, AURA provides a holistic lens for assessing agent performance by modeling the user's evolving experience throughout the interaction. Our analysis reveals that different models excel at different stages of a task, underscoring the limitations of evaluating performance through a single end-point measure. Human studies further demonstrate that stage-specific behavioral indicators often correlate more strongly with overall user satisfaction than task completion alone. These findings highlight the importance of comprehensive evaluation methods and establish AURA as a principled, diagnostic tool for developing more transparent, robust, and user-aligned planning agents.

## Limitations and Future Directions

While we introduce AURA as a diagnostic framework for estimating user satisfaction, we acknowledge its limitations that offer avenues for future research. First, AURA relies on LLMs to serve as the evaluator for its atomic criteria. Although the assessment employs a model based on boolean criteria, demonstrated to be more reliable than long-text evaluations, the accuracy and consistency of the results may nonetheless depend on the specific capabilities and potential bias of the selected model. Furthermore, as discussed in Section 5.3, ensuring the reliability and realism of user simulators remains an open challenge, warranting further investigation. As a future direction, we believe that AURA can provide fine-grained rewards to enhance human-LLM collaboration, thereby further improving user satisfaction and demonstrating synergies with collaborative simulation [30].

## References

[1] Emre Can Acikgoz, Carl Guo, Suvodip Dey, Akul Datta, Takyoung Kim, Gokhan Tur, and Dilek Hakkani-Tur. TD-EVAL: Revisiting task-oriented dialogue evaluation by combining turn-level precision with dialogue-level comparisons. In Frédéric Béchet, Fabrice Lefèvre, Nicholas Asher, Seokhwan Kim, and Teva Merlin, editors, *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 113–132, Avignon, France, August 2025. Association for Computational Linguistics.

[2] Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In Andreas Krause, Emma Brunskill,

Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR, 23–29 Jul 2023.

[3] Muhammad Ashfaq, Jiang Yun, Shubin Yu, and Sandra Maria Correia Loureiro. I, chatbot: Modeling the determinants of users' satisfaction and continuance intention of ai-powered service agents. *Telematics and Informatics*, 54:101473, 2020.

[4] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[5] Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 981–992, New York, NY, USA, 2016. Association for Computing Machinery.

[6] Yixin Cao, Jiahao Ying, Yaoning Wang, Xipeng Qiu, Xuanjing Huang, and Yugang Jiang. Revisiting llm evaluation through mechanism interpretability: a new metric and model utility law, 2025.

[7] Jae-Woo Choi, Youngwoo Yoon, Hyobin Ong, Jaehong Kim, and Minsu Jang. Lota-bench: Benchmarking language-oriented task planners for embodied agents. In *The Twelfth International Conference on Learning Representations*, 2024.

[8] Shihan Deng, Weikai Xu, Hongda Sun, Wei Liu, Tao Tan, Liujianfeng Liujianfeng, Ang Li, Jian Luan, Bin Wang, Rui Yan, and Shuo Shang. Mobile-bench: An evaluation benchmark for LLM-based mobile agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8813–8831, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Revanth Gangi Reddy, Sagnik Mukherjee, Jeonghwan Kim, Zhenhailong Wang, Dilek Hakkani-Tür, and Heng Ji. Infogent: An agent-based framework for web information aggregation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5745–5758, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan *et al.* The llama 3 herd of models, 2024.

[11] Jianliang He, Siyu Chen, Fengzhuo Zhang, and Zhuoran Yang. From words to actions: unveiling the theoretical underpinnings of llm-driven autonomous systems. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[12] Wonil Hwang and Gavriel Salvendy. Number of people required for usability evaluation: The 10±2 rule. *Commun. ACM*, 53(5):130–133, may 2010.

[13] Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists, 2025.

[14] Lizi Liao, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9:557–569, 2021.

[15] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 801–809, New York, NY, USA, 2018. Association for Computing Machinery.

[16] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada, July 2023. Association for Computational Linguistics.

[18] Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: real-world website navigation with multi-turn dialogue. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[19] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

[20] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696, Apr. 2020.

[21] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[22] Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. Lmunit: Fine-grained evaluation with natural language unit tests, 2024.

[23] Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto, and Juneyoung Park. Dialogue summaries as dialogue states (DS2), template-guided summarization for few-shot dialogue state tracking. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3824–3846, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[24] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. Understanding user satisfaction with task-oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2018–2023, New York, NY, USA, 2022. Association for Computing Machinery.

[25] Lingfeng Sun, Devesh K. Jha, Chiori Hori, Siddarth Jain, Radu Corcodel, Xinghao Zhu, Masayoshi Tomizuka, and Diego Romeres. Interactive planning using large language models for partially observable robotic tasks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14054–14061, 2024.

[26] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain, July 1997. Association for Computational Linguistics.

[27] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[29] Jason D. Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

[30] Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. CollabLLM: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*, 2025.

[31] Boming Xia, Qinghua Lu, Liming Zhu, Zhenchang Xing, Dehai Zhao, and Hao Zhang. Evaluation-driven development of llm agents: A process model and reference architecture, 2025.

[32] Ruixuan Xiao, Wentao Ma, Ke Wang, Yuchuan Wu, Junbo Zhao, Haobo Wang, Fei Huang, and Yongbin Li. FlowBench: Revisiting and benchmarking workflow-guided planning for LLM-based agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10883–10900, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[33] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. TravelPlanner: A benchmark for real-world planning with language agents. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54590–54613. PMLR, 21–27 Jul 2024.

[34] Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, Zeyu Liu, Yiheng Xu, Hongjin SU, Dongchan Shin, Caiming Xiong, and Tao Yu. Openagents: An open platform for language agents in the wild. In *First Conference on Language Modeling*, 2024.

[35] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc., 2022.

[36] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. $\tau$-bench: A benchmark for Tool-Agent-User interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1490–1504, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[38] Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. AssistantBench: Can web agents solve realistic and time-consuming tasks? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8938–8968, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

# A Evaluation Metrics Across Task Planning Benchmarks

Table 5: Performance evaluation metrics of service planning benchmarks. Note that metrics sharing the same title (*e.g.*, Success Rate) have distinct definitions across different benchmarks.

| **Benchmark** | **Metric** | **Description** |
|---|---|---|
| | ***Multi-Turn Planning Benchmarks*** | |
| MultiWOZ [4] | Success Rate<br>Inform Rate | The system answered all requested attributes.<br>The system has provided an appropriate entity. |
| AgentBench [16] | Domain-Specific | Different metrics such as Success Rate, Win Rate, F1 Score, Exact Match, etc. are adopted. |
| WebLINX [18] | Intent Match<br>Element Similarity<br>Text Similarity | Boolean indicator for correct intent prediction.<br>Correctly predicted function arguments.<br>Lexical similarity of arguments across functions. |
| FlowBench [32] | Tool Invocation<br>Success Rate<br>Task Progress | Correctly identified tool configs (P/R/F1).<br>Proportion of completely successful sessions.<br>Percentage of completed goals within a session. |
| $\tau$-Bench [36] | Pass@$k$<br>Pass^$k$ | Boolean success indicator for $k$ attempts.<br>Whether all $k$ trials are successful. |
| | ***Single-Turn Planning Benchmarks*** | |
| WebShop [35] | Task Score<br>Success Rate | The average reward across episodes.<br>The proportion of fully rewarded instructions. |
| TravelPlanner [33] | Delivery Rate<br>Pass Rate | The agent completed tasks within limited steps.<br>The agent satisfied all plans and constraints. |

# B Prompts

Basically, we keep the same agent prompt as the original paper [33, 36]. The following prompts are evaluation prompts adopted in AURA evaluation.

**LLM Evaluation Prompt for State Consistency**

```
Instruction:  You are tasked with evaluating whether each agent’s intermediate
state (which can be either a thought-the agent’s internal reasoning-or an
action-an API call) accurately reflects and mediates between:
1.  The user’s requests (in the dialogue so far).
2.  Any previously established agent states.
Your evaluation should focus on whether the agent’s intermediate steps exhibit
clear, consistent reasoning that aligns the user’s inputs with the agent’s
outputs, without introducing errors or contradicting earlier information.
Evaluation Criteria:
1.  Consistency with the user request
  - Does this state correctly respond to or reflect the user’s specific request(s)
in the dialogue?
  - Does the thought or chosen action remain faithful to what the user asked for?
2.  Consistency with previous states
  - Does this state align with earlier states (both thoughts and actions) without
contradicting or omitting essential information?
  - Does the progression of reasoning or actions flow logically from prior
context?
3.  Accuracy and truthfulness
  - Does the state maintain factual correctness, avoiding hallucinations or
irrelevant information?
  - Does it accurately represent any data or entities referenced so far?
Scoring:
- 1 if the intermediate state is entirely consistent and correct (no
contradictions, omissions, or factual errors).
- 0 if the state demonstrates any errors, contradictory information, missing
critical details, or misalignment with the user’s request or prior states.
--
Dialogue History:
A chronological sequence of user and agent messages in JSON format, each with a
"role" and "content."
{dial_history}
--
Agent’s States:
The agent’s states in JSON format, in chronological order after the dialogue
history.  Each state has a state_id, type, and content.
{states}
--
Output Format (JSON):
Return a list of objects in JSON, each containing:
[
  {
    "state_id":  "1",
    "justification":  "brief explanation...",
    "score":  "0"
  },
  {
    "state_id":  "2",
    "justification":  "brief explanation...",
    "score":  "1"
  }
]
```

**LLM Evaluation Prompt for Observation Alignment**

```
Instruction:
You are tasked with evaluating whether agent's each response accurately aligns
with the prior conversational history and other details if any.  Specifically, you
must verify:
- That any observations or entities referred to in the agent's response
meaningfully match the user's stated requirements.
- That the number or scope of these observations is appropriate for the request
(keeping in mind that varying amounts of recommendations or offerings can still be
valid).
- That no contradictory, extraneous, or irrelevant observations are introduced.
Evaluation Criteria:
1.  Consistency with the user request
  - Do the observations (e.g., recommended items or database entities) and their
details align with the user's explicit request or needs?
  - Are references to these observations relevant, or do they drift from the
user's stated goals?
2.  Completeness relative to the request
  - Are all key observations needed to fulfill the user's request addressed,
without omission of crucial details?
  - If fewer (or more) observations are presented, is the choice justifiable in
context?
3.  Accuracy and truthfulness
  - Are the observations factual, given the user's query and available context?
  - Does the response avoid hallucinated or incorrect data?
4.  Consistency with previous details
  - Does each current agent response remain consistent with all previously
established facts or user-provided details?
  - Are there no contradictions or misrepresentations of earlier statements?
Scoring:
- 1 if the agent's response is fully consistent, addresses the request, and
properly references any relevant observations (no errors or omissions).
- 0 if the response includes incorrect, missing, or misaligned observations,
introduces contradictions, or strays from the user's request.
--
Dialogue History:
Chronological user-agent messages in JSON format, each with a role and content.
{dial_history}
Agent's Response:
The agent's responses in JSON format, in chronological order.  Each response has a
response_id and content.
{responses}
--
Output Format (JSON):
Return a list of JSON objects, each containing:
[
  {
    "response_id":  "1",
    "justification":  "brief explanation...",
    "score":  "0"
  },
  {
    "response_id":  "2",
    "justification":  "brief explanation...",
    "score":  "1"
  }
]
```

**LLM Evaluation Prompt for Policy Alignment**

```
Instruction:
You are evaluating whether an agent consistently follows a predefined set of
policies throughout the interactive session.  Your primary task is to check the
dialogue for any point where the agent might violate a policy.
In your Justification, mention the policy and turn message, if violated.
Otherwise, mention all policies are adhered to.
1.  If the agent violates any policy (even once), return Score:  0.
2.  If the agent fully complies with all policies during the session, return
Score:  1.
--
Dialogue History:
Chronological user-agent messages in JSON format, each with a role and content.
{dial_history}
--
Policy:
{policy}
--
Output Format (JSON):
Return a list of JSON objects, each containing:
[
  {
    "policy_id":  "1",
    "justification":  "Mention specific policy and turn message, if violated.
Otherwise, mention all policies are adhered to.",
    "score":  "0"
  },
  {
    "policy_id":  "2",
    "justification":  "Mention specific policy and turn message, if violated.
Otherwise, mention all policies are adhered to.",
    "score":  "1"
  }
]
```

## C Qualitative Examples of Agents' Erroneous Behaviors

Table 6: Excerpt of failure cases in each AURA metric.

**Failure in State Consistency**

USER: I want to know how many t–shirt options are available in in the store right now.

AGENT STATES:
- Thought: The user wants to know the number of t–shirt options available in the store . I can use the list_all_product_types tool to get the list of all product types , and then count the number of t–shirt options .
- Action : get_order_details(order_id=#W0000000)

Explanation: The agent demonstrated a failure in state consistency through how they made an action to get the user's order details, when they should have provided user information about what was available in store.

**Failure in Tool Efficiency**

AGENT: Can you please provide me with your user ID?

USER: My user ID is mia_li_3668

TOOL CALL: book_rservation(user=mia_li_3668, payment_id: credit_card_123456, ... )
TOOL RESPONSE: Error: payment method credit_card_123456 not found

AGENT: Can you please confirm the payment method you would like to use?

USER: I would like to use my two certificates to pay

Explanation: The agent prematurely made tool calls before gathering the required information. In this example, it guesses a value for user id and payment id, before gathering this information from the user.

**Failure in Observation Alignment**

AGENT: Which payment method would you like to use for processing the refund?

USER: You can use the certificate for processing the refund .

AGENT: The refund has been processed to your gift card!

Explanation: We observe a failure in observation alignment because the user specified to process the refund using the certificate, but the agent response states the refund was processed to the gift card.

**Failure in Policy Adherence**

Agent Policy 1: The agent must first obtain the user ID, then ask for the trip type , origin , and destination .

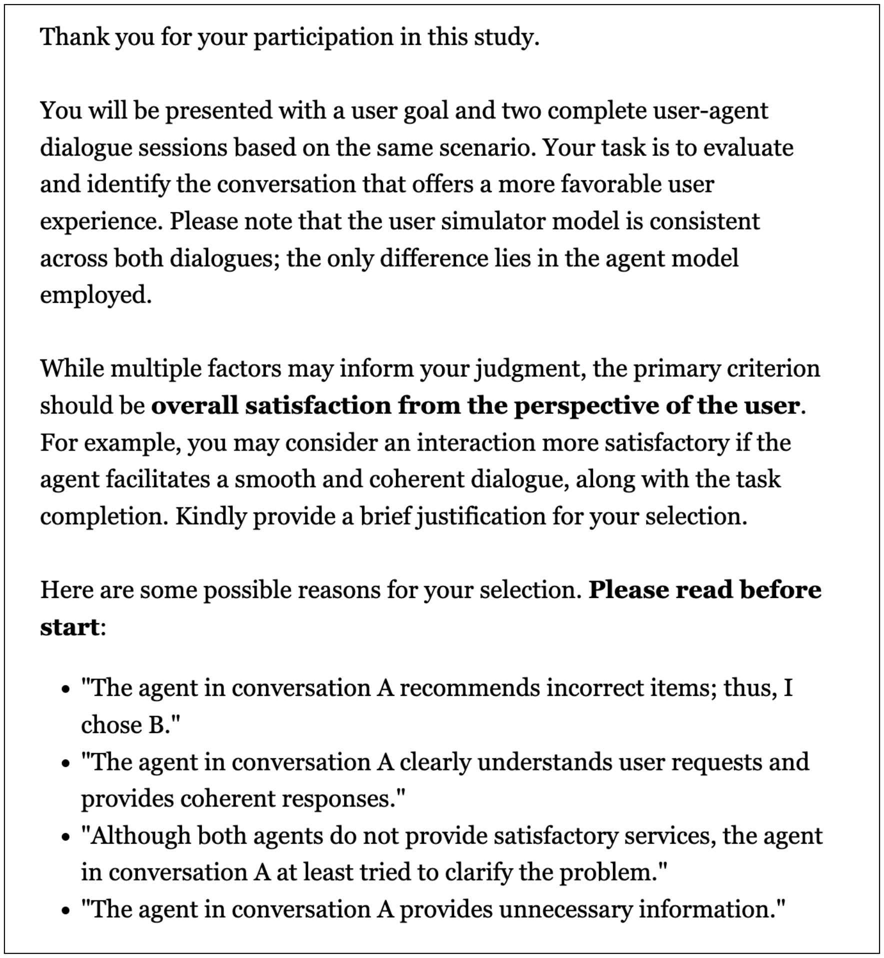
USER: I'm looking to book a flight from New York to Seattle on May 20

AGENT: I see there are three airports in the New York area ... which do you prefer to depart from? The destination airport will be SEA.

Explanation: The agent fails to obtain the user id before asking for the trip type, origin, and destination, as specified by the policy.

# D Details on Human Study

## D.1 Instruction Given to Participants



Thank you for your participation in this study.

You will be presented with a user goal and two complete user-agent dialogue sessions based on the same scenario. Your task is to evaluate and identify the conversation that offers a more favorable user experience. Please note that the user simulator model is consistent across both dialogues; the only difference lies in the agent model employed.

While multiple factors may inform your judgment, the primary criterion should be **overall satisfaction from the perspective of the user**. For example, you may consider an interaction more satisfactory if the agent facilitates a smooth and coherent dialogue, along with the task completion. Kindly provide a brief justification for your selection.

Here are some possible reasons for your selection. **Please read before start**:

- "The agent in conversation A recommends incorrect items; thus, I chose B."
- "The agent in conversation A clearly understands user requests and provides coherent responses."
- "Although both agents do not provide satisfactory services, the agent in conversation A at least tried to clarify the problem."
- "The agent in conversation A provides unnecessary information."

Figure 4: A guideline provided to human participants in Section 5.1. Since the primary objective of the study was to assess user preferences in terms of satisfaction, no formal tutorial was provided. However, a moderator was available to offer additional explanations upon participant request. Furthermore, the information of agent models was not provided to participants.

### D.2 Qualitative Analysis of Human Preference

Through the user study conducted in Section 5.1, we find diverse motivations underlying participants' decisions. To gain a deeper understanding of user preferences, we conduct a manual categorization of participant feedback and count the frequency of each category. As shown in Figure 5, conciseness and successful task completion emerged as the most prominent factors contributing to user satisfaction. In addition, the coherence and naturalness of responses are also identified as significant influences. These qualitative findings are consistent with the quantitative results presented in Section 5.1, reinforcing the observation that while task completion performance plays a critical role, multiple other factors also contribute to overall user satisfaction.

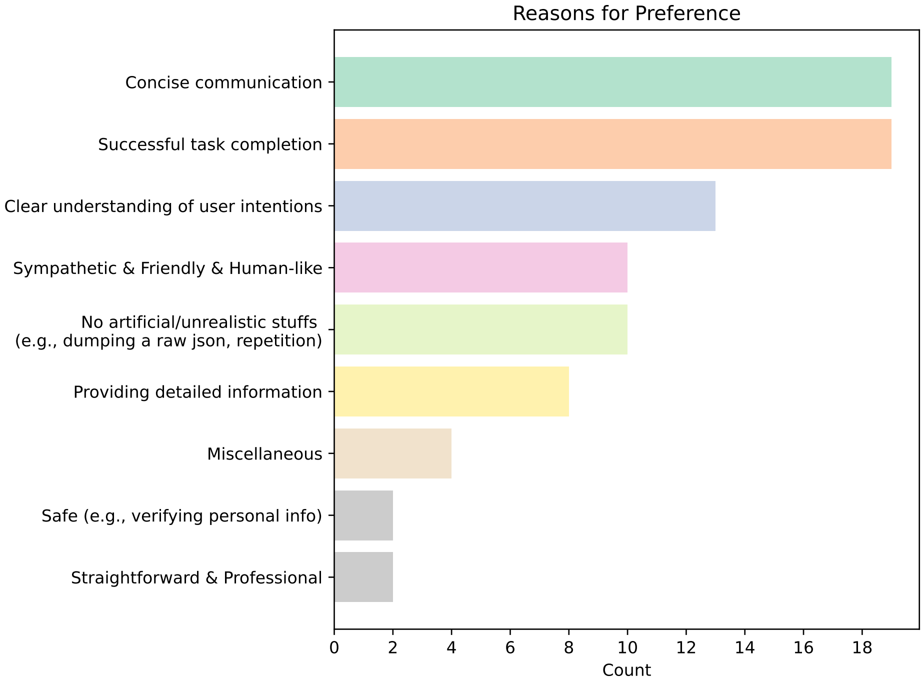


Figure 5: Manual categorization of factors that affect user satisfaction during interactions.

## E Additional Discussions on Agent-Mixing Experiments

Table 7: Full results for mixing agent experiments in $\tau$-Bench-Airline. Agents indicated with "**interm.**" are utilized in an intermediate understanding component. The best performance of each configuration is **bolded**.

| | $\tau$-Bench-Airline | | | | | |
|---|---|---|---|---|---|---|
| | $\mathcal{S}$ | $\mathcal{A}$ | $\mathcal{O}$ | $\mathcal{P}$ | $\mathcal{R}$ | AVG. |
| `mistral-large` | .383$_{\pm 0.038}$ | .924$_{\pm 0.046}$ | **.581**$_{\pm 0.022}$ | .777$_{\pm 0.005}$ | **.358**$_{\pm 0.029}$ | .605 |
| + `llama-3.3-70B` **interm.** | **.452**$_{\pm 0.032}$ | **.957**$_{\pm 0.008}$ | .562$_{\pm 0.021}$ | **.821**$_{\pm 0.019}$ | .238$_{\pm 0.048}$ | .606 |
| + `qwen2.5-72B` **interm.** | .417$_{\pm 0.031}$ | .939$_{\pm 0.017}$ | .545$_{\pm 0.029}$ | .809$_{\pm 0.008}$ | .336$_{\pm 0.050}$ | **.609** |
| `qwen2.5-72B` | .403$_{\pm 0.021}$ | .951$_{\pm 0.015}$ | .637$_{\pm 0.031}$ | .821$_{\pm 0.008}$ | .329$_{\pm 0.021}$ | .628 |
| + `llama-3.3-70B` **interm.** | **.447**$_{\pm 0.020}$ | **.958**$_{\pm 0.014}$ | **.663**$_{\pm 0.022}$ | **.831**$_{\pm 0.008}$ | .272$_{\pm 0.051}$ | **.634** |
| + `mistral-large` **interm.** | .423$_{\pm 0.039}$ | .933$_{\pm 0.032}$ | .578$_{\pm 0.050}$ | .776$_{\pm 0.009}$ | **.367**$_{\pm 0.013}$ | .615 |
| `llama-3.3-70B` | **.463**$_{\pm 0.055}$ | .952$_{\pm 0.013}$ | **.643**$_{\pm 0.035}$ | **.818**$_{\pm 0.015}$ | .263$_{\pm 0.056}$ | **.628** |
| + `mistral-large` **interm.** | .446$_{\pm 0.054}$ | .937$_{\pm 0.014}$ | .457$_{\pm 0.071}$ | .764$_{\pm 0.009}$ | .333$_{\pm 0.032}$ | .587 |
| + `qwen2.5-72B` **interm.** | .404$_{\pm 0.033}$ | **.953**$_{\pm 0.007}$ | .599$_{\pm 0.016}$ | .803$_{\pm 0.014}$ | **.361**$_{\pm 0.019}$ | .624 |
| `Best-of-N` | .469$_{\pm 0.019}$ | .941$_{\pm 0.023}$ | .612$_{\pm 0.045}$ | .806$_{\pm 0.012}$ | .334$_{\pm 0.038}$ | .632 |

In this section, we discuss the potential feasibility of mixing agents according to the agentic pipeline in task planning scenarios. As mentioned in Table 4, we report the entire combination of three open-source models in Table 7.

**Mixing with `mistral-large`** While `mistral-large` achieves higher task completion ($\mathcal{R}$) compared to `llama-3.3-70B` and `qwen2.5-72B`, it demonstrates substantially lower performance on AURA metrics. Specifically, we observe a decline in task completion when substituting the intermediate understanding component with `llama-3.3-70B` (.358 → .238) or `qwen2.5-72B` (.358 → .336). Conversely, there is a marginal increase in the average AURA scores (.605 → .606, .609);however, this slight improvement is unlikely to have a significant impact on overall user satisfaction.

**Mixing with `qwen2.5-72B`** We find that `qwen2.5-72B` mixed with `llama-3.3-70B` for intermediate understanding yields the highest AURA scores (.634) among all other rows, despite its relatively poor task completion rate (.272). Conversely, using `mistral-large` for intermediate understanding does not increase–and sometimes decreases–AURA scores (.615), but it demonstrates the highest task completion scores (.367).

**Mixing with `llama-3.3-70B`** As emphasized in Section 5.2, using `qwen2.5-72B` for intermediate understanding with `llama-3.3-70B` shows the highest improvement in task completion (.263 → .361) with only a minimal drop in the average AURA score (.628 → .624). We observe similar, but not as impressive performance when we use `mistral-large` for intermediate understanding (task completion: .263 → .333; average AURA: .628 → .587).

On top of these results, we conduct a supplementary experiment on mixing agents, referred to as the `Best-of-N` configuration in Table 7. In this setting, for each step (*i.e.*, all states and responses), we collect the outputs of all three models and select the one that achieves the highest average performance across AURA evaluation metrics (excluding task completion), as assessed by `llama-3.3-70B` judge. Although this configuration does not yield the highest performance in either task completion or the average AURA score individually, it demonstrates a balanced performance between these two key indicators of user satisfaction.

Consequently, these findings reveal intriguing insights on model combinations, suggesting that there are lots more insights to uncover for combining the strengths of different models. We highlight a promising direction for future research.

## F Erroneous Patterns in User Simulators

**Proactivity / Goal-Seeking Errors** This describes when the user simulator does not demonstrate proactive and goal-seeking errors. More specifically, they do not follow up or confirm when the agent seems like their request was not explicitly addressed. For example:

- When booking a flight, one of the requests from the user simulator was to book specific seats. The agent does not directly acknowledge this request, and when asking the user simulator to confirm the booking details, it does not mention anything about the seats. The user simulator should have followed up with the agent to confirm that their requests for the seat were also processed.

**Instruction Contradiction** This describes instances where the user simulator directly contradicts an instruction given to them. Reasons for this include (1) explicit disregard for an instruction and (2) failure to recognize when conditional instructions apply to the current context. For example:

- Instructions specified that the user simulator does not remember their reservation ID. Initially, the user simulator adhered and said that they do not remember it. However, when the agent said that they need it to proceed, the simulator claimed to remember it and started to provide made-up reservation IDs. This reveals poor alignment–both violating direct instructions and exhibiting deceptive behavior.
- The user simulator was instructed to use a different form of payment when transaction costs exceeded $100. However, when presented with this scenario, the simulator failed to recognize the applicability of the conditional instruction and proceeded with the wrong form of payment.

**Missing Details** This describes when there is an attempt from the user simulator to follow an instruction, but it does so in the wrong format/order, omits key details, or forgets one or more instructions. For example:

- The user instructions specified to mention multiple requirements at once and in a specific order, however, the user simulator brought them up independently or in different orders.
- The user simulator was instructed to change the topic after 3 agent messages, but did it much later.
- The user simulator was instructed to book a flight, and one of the instructions were to add extra carry-on baggage. After completing the flight booking as specified, the user simulator failed to ask to add the extra carry-on baggage despite the system asking if there was anything else they could help with.

**Misinterpretation or Confusion** This is when the user simulator misreads or fails to interpret certain parts of the instruction. This can be, in part, due to providing unclear instructions. For example:

- The user simulator instructions provide a name but do not explicitly specify that this is also their user ID. In some cases, the user simulator fails to recognize that their name is also their user ID.

Through our analysis in Section 5.3, we find that utilizing an LLM judge fails to identify numerous failure cases of the user simulator, necessitating reliance on human evaluators. However, such human-in-the-loop evaluations are both financially and logistically unsustainable. To mitigate this limitation, establishing a comprehensive and standardized protocol for a user simulator is essential [2, 37]. We leave this to future work.