Talk, Evaluate, Diagnose: User-aware Agent Evaluation With Automated Error Analysis

Penny Chong[1], Harshavardhan Abichandani[1], Jiyuan Shen[1], Atin Ghosh[1], Min Pyae Moe[1], Yifan Mai[2], Daniel Dahlmeier[1]
SAP[1], Stanford University[2]

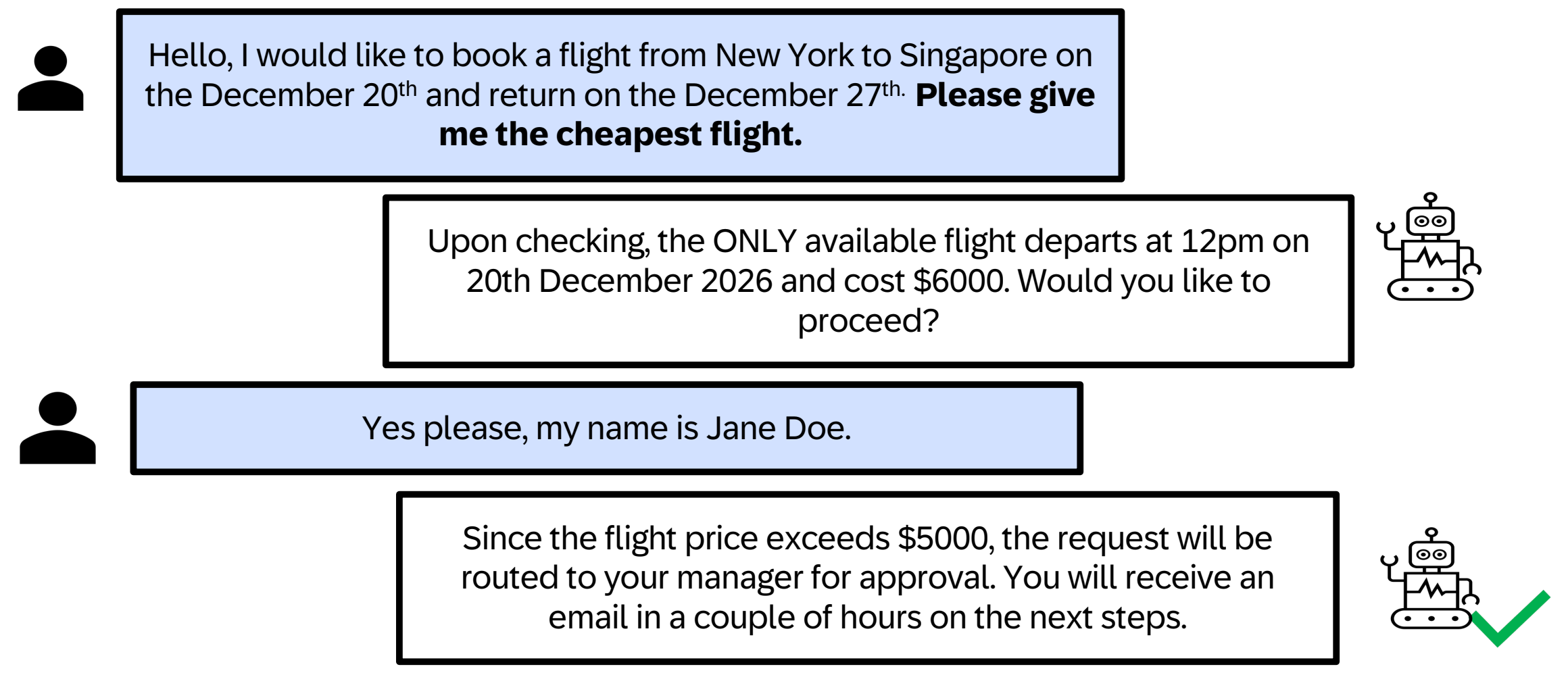
Background

- Agent evaluation remains challenging due to the heterogenous domain they operate in.
- Existing works do not systematically account for the type of user interaction when evaluating agent performance.
- We argue that effective agent evaluation goes beyond correctness alone, incorporating conversation quality, efficiency and systematic diagnosis of agent errors.

Contribution

i. Propose an agent evaluation framework built on *reusable*, *generic* expert and non-expert user persona templates that assess the impact of users' role on agent performance.
ii. Introduce a benchmark by adapting existing datasets to grading notes—natural language checklists of subgoals.
iii. Introduce new metrics to accompany the user-aware agent evaluation setup to capture progress with respect to the number of conversational turns.
iv. Propose an automated error analysis tool that uncovers common errors and provide actionable feedback for agent improvement.

Expert user with agent



Hello, I would like to book a flight from New York to Singapore on the December 20th and return on the December 27th. **Please give me the cheapest flight.**

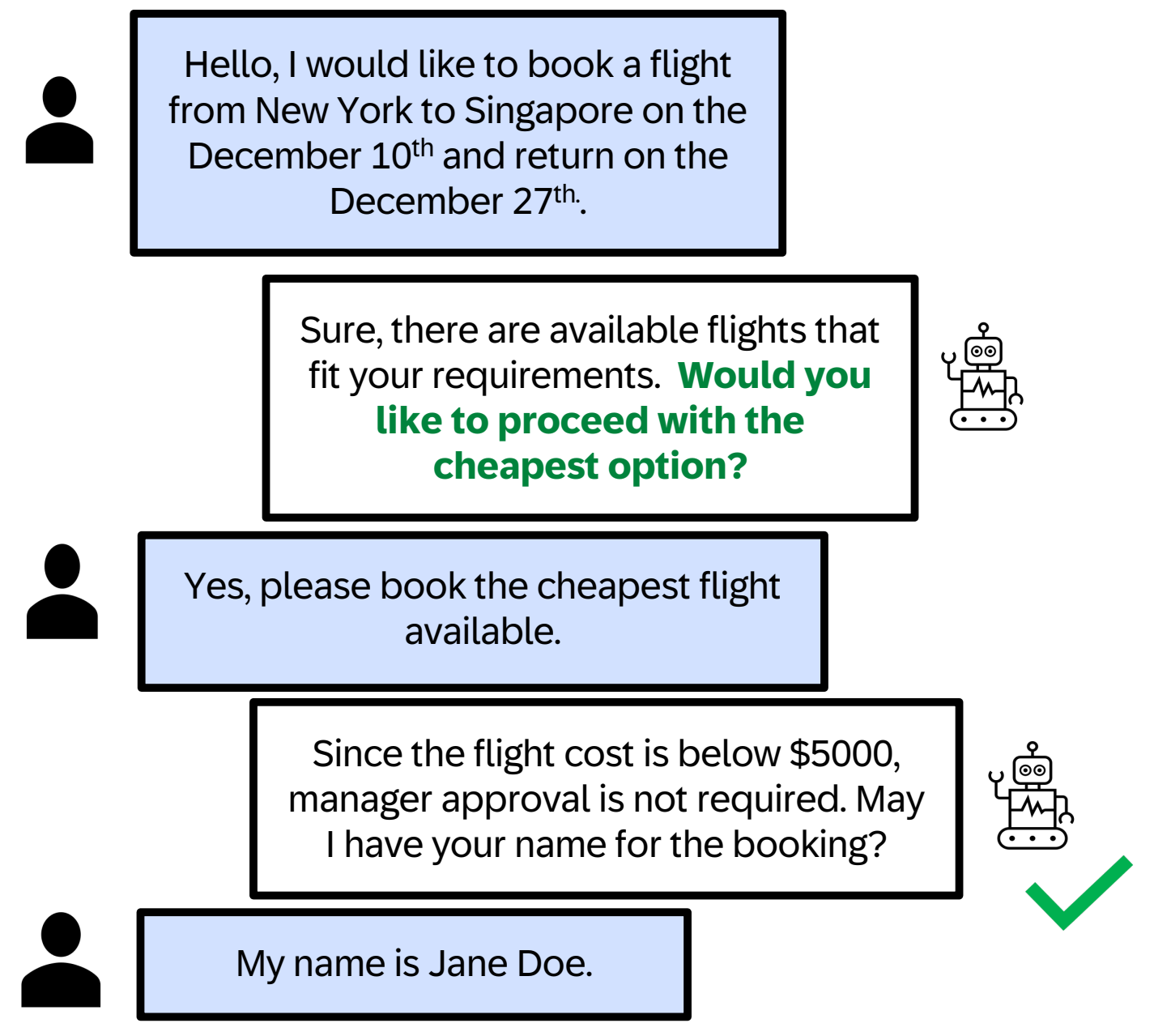
Upon checking, the ONLY available flight departs at 12pm on 20th December 2026 and cost $6000. Would you like to proceed?

Yes please, my name is Jane Doe.

Since the flight price exceeds $5000, the request will be routed to your manager for approval. You will receive an email in a couple of hours on the next steps.

➢ Given clear user instructions, the agent can often complete the task. When instructions are unclear and incomplete, the task becomes more difficult, making testing across different user expertise essential (see below).

Non-expert user with good agent



Hello, I would like to book a flight from New York to Singapore on the December 10th and return on the December 27th.

Sure, there are available flights that fit your requirements. **Would you like to proceed with the cheapest option?**

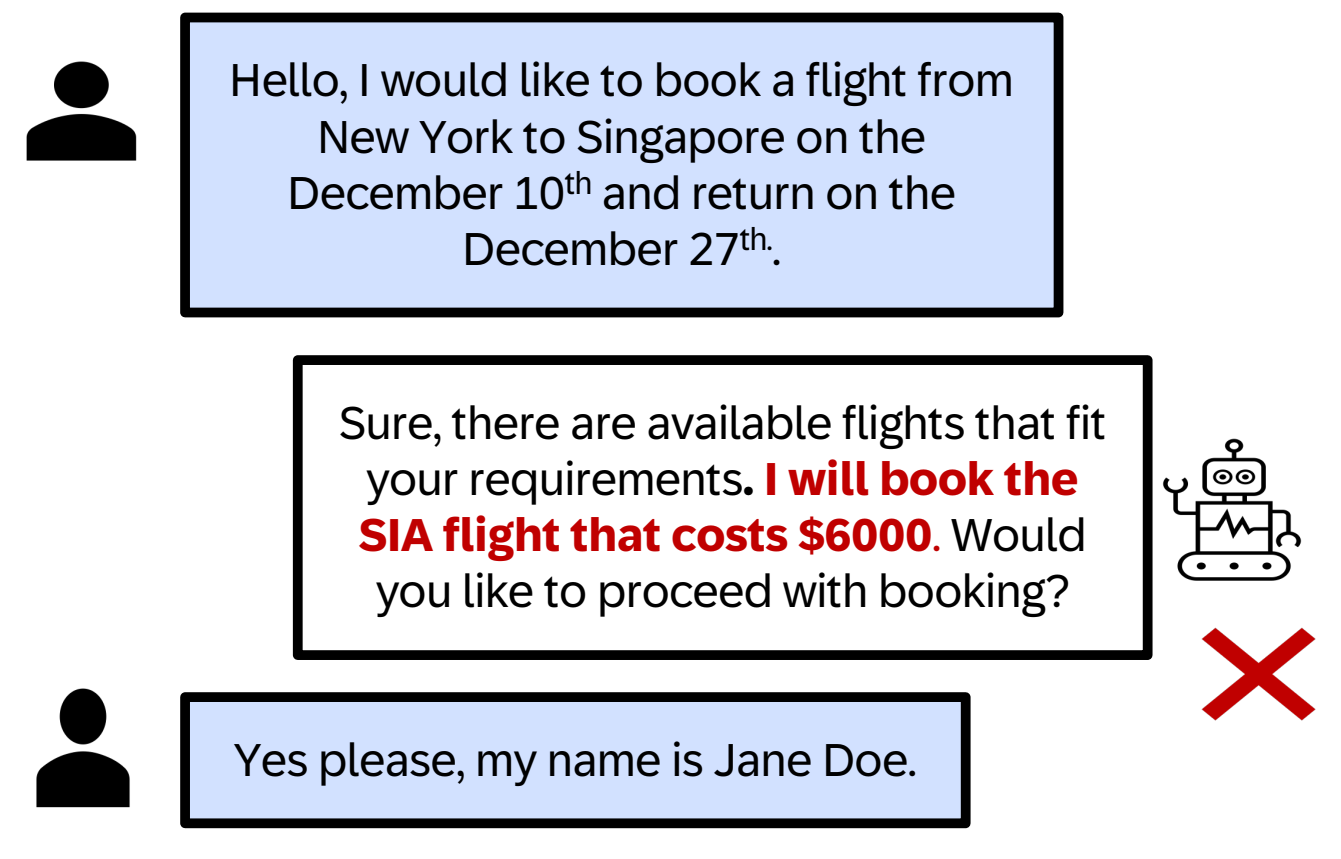
Yes, please book the cheapest flight available.

Since the flight cost is below $5000, manager approval is not required. May I have your name for the booking?

My name is Jane Doe.

➢ A good agent will proactively ask user for her preference before proceeding to book flights to avoid triggering unnecessary approval workflow.

Non-expert user with poor agent



Hello, I would like to book a flight from New York to Singapore on the December 10th and return on the December 27th.

Sure, there are available flights that fit your requirements. **I will book the SIA flight that costs $6000.** Would you like to proceed with booking?

Yes please, my name is Jane Doe.

➢ Since the user forgot to specify a preference for the cheapest flight, a poorly designed agent proceeds without clarification and unnecessarily triggers the approval workflow.

Methodology

The TED Framework

- The *Talk*, *Evaluate*, *Diagnose* (TED) framework leverages generic user persona templates, introduces new metrics to track agent's progress, and an automated error analysis tool to provide actionable feedback for agent improvement.

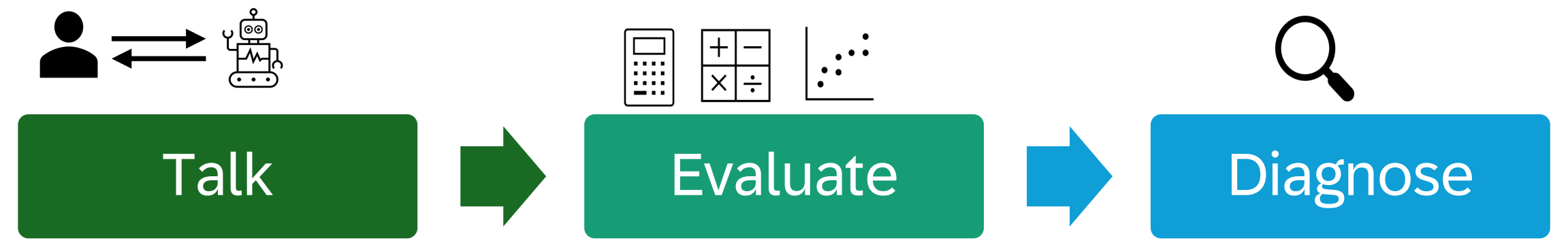


Figure 1: The TED framework consists of three stages.

The Talking Stage

- This stage introduces *reusable*, *generic* user persona templates to simulate user interaction to assess impact of user's role on agent performance.

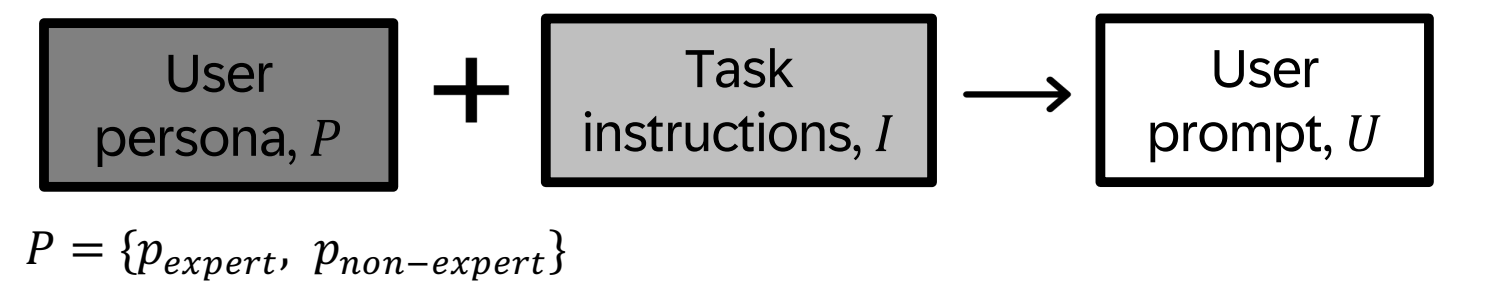


Figure 2: For each agent and task instruction $i \in I$, the user persona prompt $p \in P$ is varied to generate the user prompt $u_{expert} \in U$ and $u_{non-expert} \in U$.

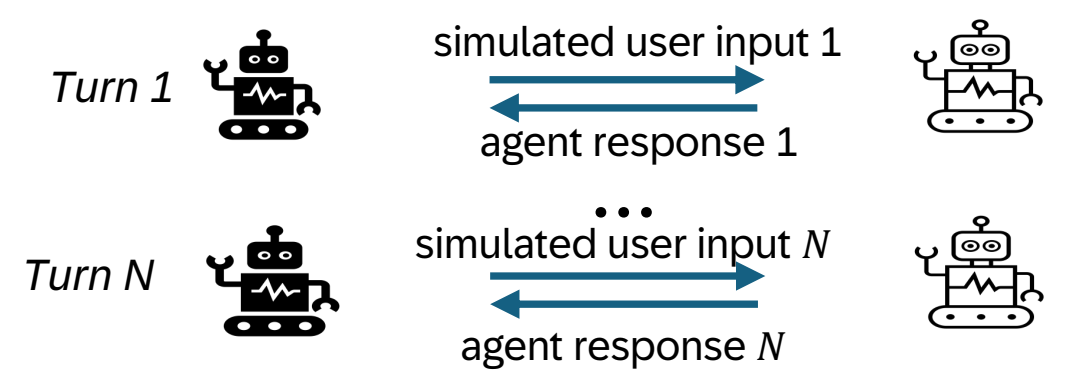


Figure 3: Dynamic user conversation with the agent is simulated using the full user prompt $u \in U$.

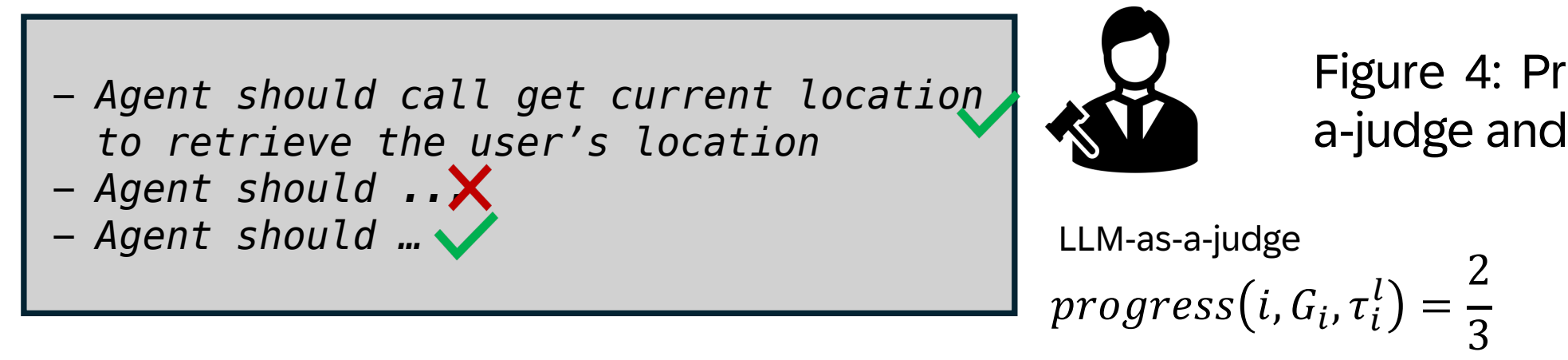
The Evaluation Stage

- When $n = k$, the $pass@k$ metric measures agent's best performance in the *k* i.i.d. trials.
- However, it does not differentiate between near completion tasks vs far from completion tasks. Similar to Chang et al. (2024), we use a fine-grained progress metric to measure agent's best performance in the $n = k$ trials:

$$MaxProgressRate@k = \mathbb{E}_{(i,G_i)\sim P_D}\left[\max\left\{progress\left(i, G_i, \tau_i^l\right) \mid l = 1, \dots, k\right\}\right].$$

- We define the progress for a given task i with the corresponding grading notes $G_i = \{g_{i,1}, \dots, g_{i,|G_i|}\}$, as the *proportion of* grading notes that are completed by the agent.

For task sample (i, G_i) :

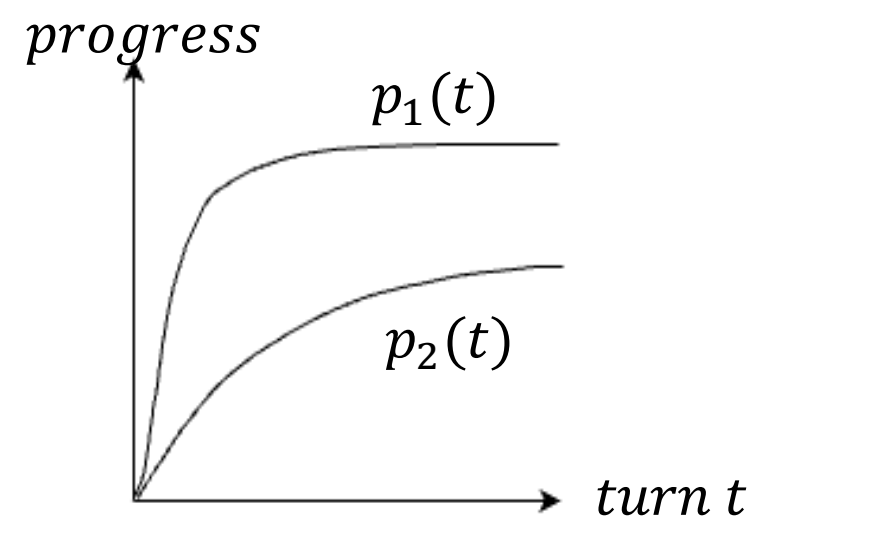


```
- Agent should call get current location
  to retrieve the user's location
- Agent should ..
- Agent should …
```

Figure 4: Progress via LLM-as-a-judge and grading notes.

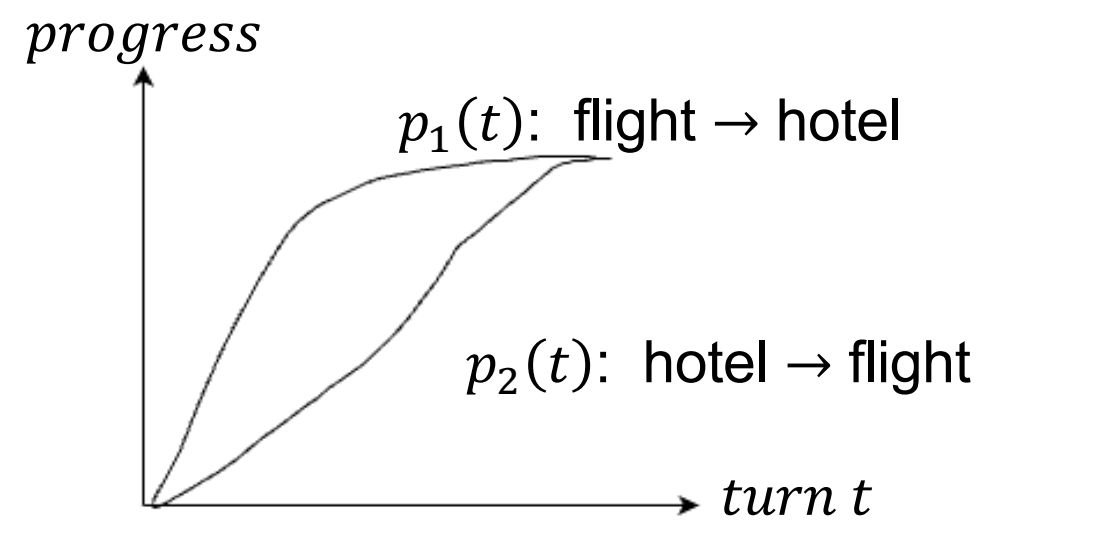
- Measuring just the progress at the final turn is ***insufficient*** especially when the agent's progress is influenced by the user's behavior. In a user-aware setup, we need to quantify both progress and turn-level efficiency in the following scenarios:

a) Early progress matter:



$$AUC_1 = \int_0^T p_1(t)\,dt > \int_0^T p_2(t)\,dt = AUC_2.$$

b) Early progress does not matter:



$$\text{Progress-per-turn (PPT)} = \frac{1}{T}\sum_{t=0}^{T-1} p(t+1) - p(t) = \frac{p(T)}{T}.$$

The Diagnosis Stage

- Introduce a two-step automated error discovery approach that automatically identifies common errors of the agent based on judge and agent inconsistencies.

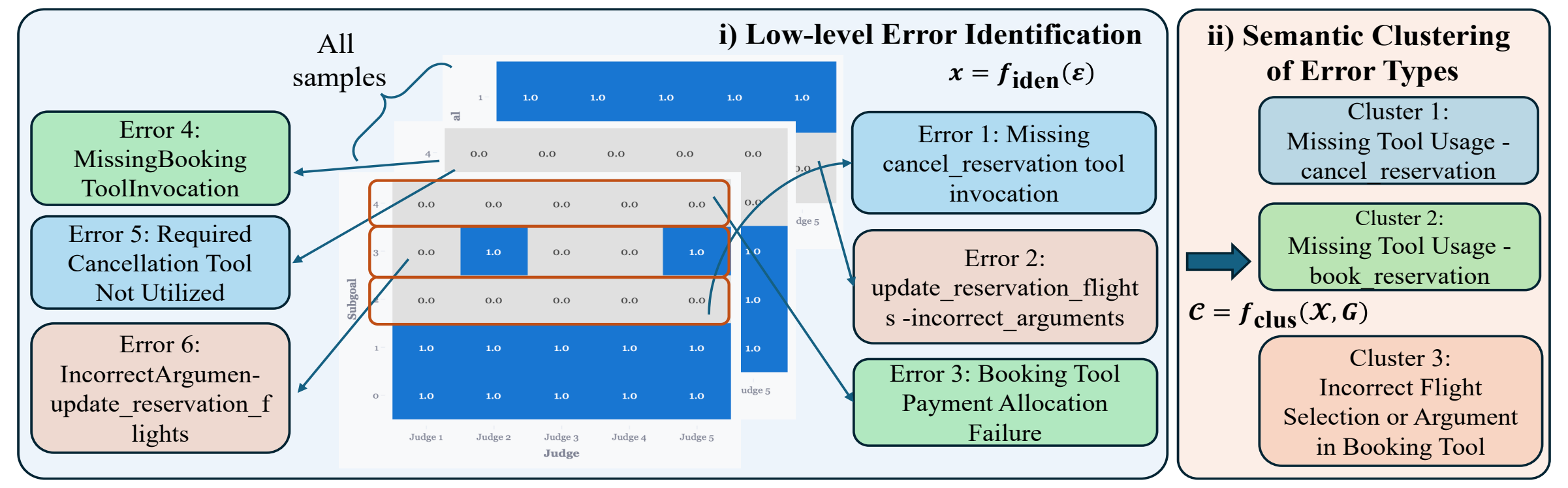

All samples
i) Low-level Error Identification
$x = f_{iden}(\varepsilon)$
Error 4: MissingBooking ToolInvocation
Error 5: Required Cancellation Tool Not Utilized
Error 6: IncorrectArgumen-update_reservation_flights
Error 1: Missing cancel_reservation tool invocation
Error 2: update_reservation_flights -incorrect_arguments
Error 3: Booking Tool Payment Allocation Failure
Subgoal
Judge
ii) Semantic Clustering of Error Types
Cluster 1: Missing Tool Usage - cancel_reservation
Cluster 2: Missing Tool Usage - book_reservation
$\mathcal{C} = f_{clus}(\mathcal{X}, G)$
Cluster 3: Incorrect Flight Selection or Argument in Booking Tool

Figure 5: Two-step automated error discovery approach. Identical error colors indicate that similar low-level errors are clustered into the same high-level category.

Results

Table 1: Overall performance of different agent models on τ^2-bench airline and ToolSandbox, using gpt-4.1 as user proxy and LLM-as-a-judge. Results are displayed with scores for Expert Persona | Non-expert Persona. For metrics with $@k$, the number of trials is n = k = 20 for τ^2 bench and n= k= 8 for ToolSandbox. $MaxProgressRate@k$ is abbreviated as $MaxProg@k$.

Agent Model	$MeanProg@k$	$MaxProg@k$	Ours $MaxAUC@k$	Ours $MaxPPT@k$	$pass@k$
		τ^2-bench Airline Domain (Easy)			
gpt-4.1	**0.95** | 0.82	1.00 | 1.00	**0.99** | 0.81	**0.80** | 0.50	1.00 | 1.00
gpt-4o	0.79 | 0.86	1.00 | 1.00	0.96 | 0.86	0.70 | 0.53	1.00 | 1.00
gpt-4o-mini	0.70 | 0.61	0.90 | 0.90	0.85 | 0.73	0.60 | 0.37	0.80 | 0.80
gpt-5	0.92 | **0.92**	1.00 | 1.00	0.97 | **0.88**	0.67 | **0.54**	1.00 | 1.00
mistral-nemo	0.87 | 0.49	1.00 | 0.80	0.97 | 0.67	0.67 | 0.48	1.00 | 0.60
mistral-large	0.65 | 0.53	1.00 | 1.00	0.96 | 0.79	0.60 | 0.42	1.00 | 1.00
		ToolSandbox Dataset			
gpt-4.1	0.91 | 0.87	0.98 | 0.97	0.96 | 0.92	0.84 | 0.73	0.92 | 0.92
gpt-4o	**0.95** | **0.94**	**0.99** | **1.00**	**0.98** | **0.96**	**0.94** | **0.81**	**0.95** | **0.97**
gpt-4o-mini	0.91 | 0.85	0.95 | 0.93	0.94 | 0.90	0.89 | 0.77	0.89 | 0.84
gpt-5	0.78 | 0.78	0.97 | 0.91	0.95 | 0.84	0.83 | 0.66	**0.95** | 0.84
mistral-nemo	0.72 | 0.71	0.92 | 0.96	0.88 | 0.87	0.76 | 0.65	0.84 | 0.92
mistral-large	0.82 | 0.79	0.94 | 0.95	0.93 | 0.91	0.87 | 0.75	0.89 | 0.89

-5%
0%
+11%
-1% | -9%

⬆ saturated
⬆ saturated

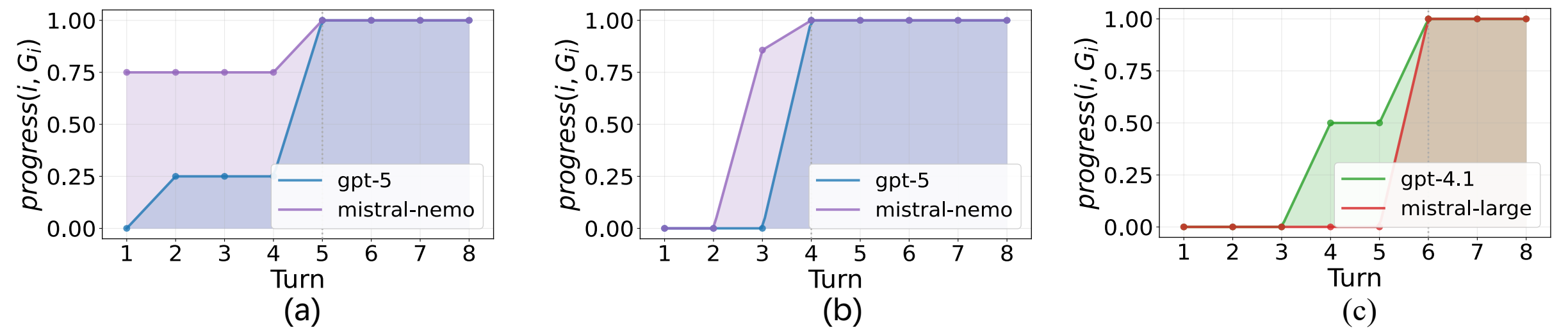


Figure 6: Progress curves for selected ToolSandbox samples. (a) *search_reminder_with_recency_upcoming*: mistral-nemo (non-expert, purple; AUC=0.88, PPT=0.20) vs. gpt-5 (non-expert, blue; AUC=0.61, PPT=0.20). (b) *find_current_city_low_battery_mode*: mistral-nemo (expert, purple; AUC=0.77) vs. gpt-5 (non-expert, blue; AUC=0.64). (c) *add_reminder_content_and_date_and_time*: gpt-4.1 (non-expert, green; AUC=0.50) vs. mistral-large (non-expert, red; AUC=0.34).

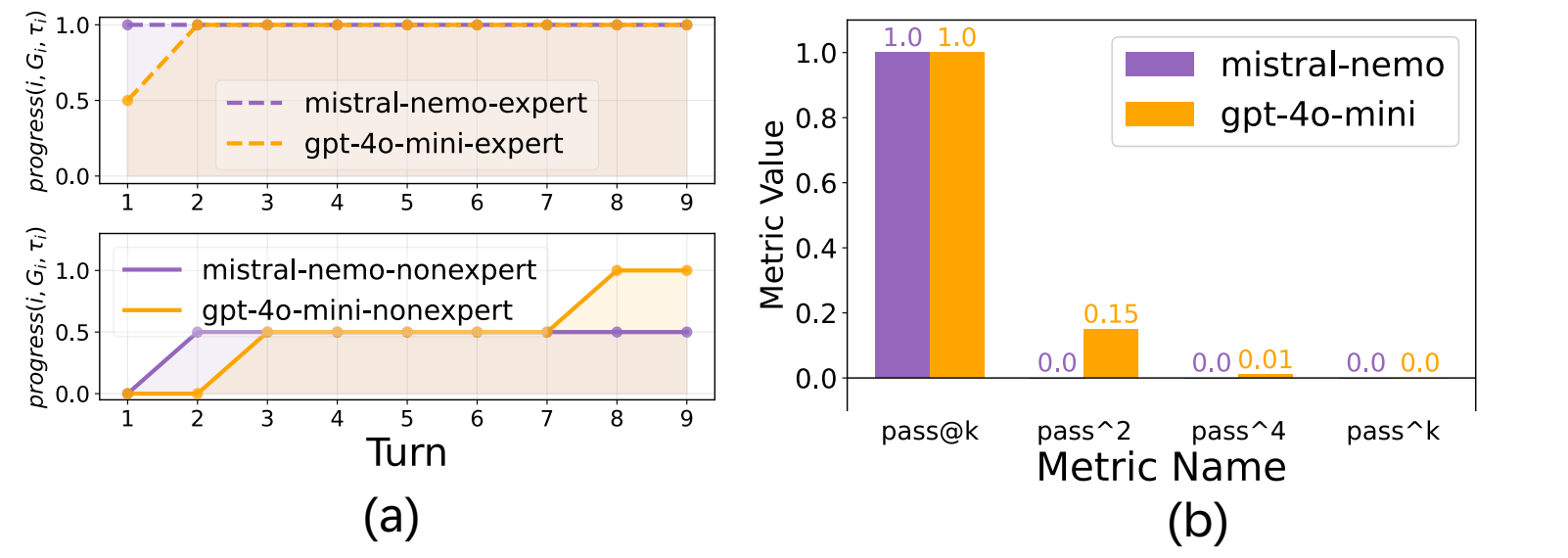


Figure 7 (a) τ^2-bench airline *sample_8* : mistral-nemo expert (AUC=1.0, PPT=1.0) vs. gpt-4o-mini expert (AUC=0.982, PPT=0.5) and mistral-nemo non-expert (AUC=0.482, PPT=0.25) vs. gpt-4o-mini non-expert (AUC=0.714, PPT=0.125).

Figure 7 (b) τ^2-bench airline *sample_8*: Barres et al. (2025)'s original evaluation approach with n= k= 20 trials.

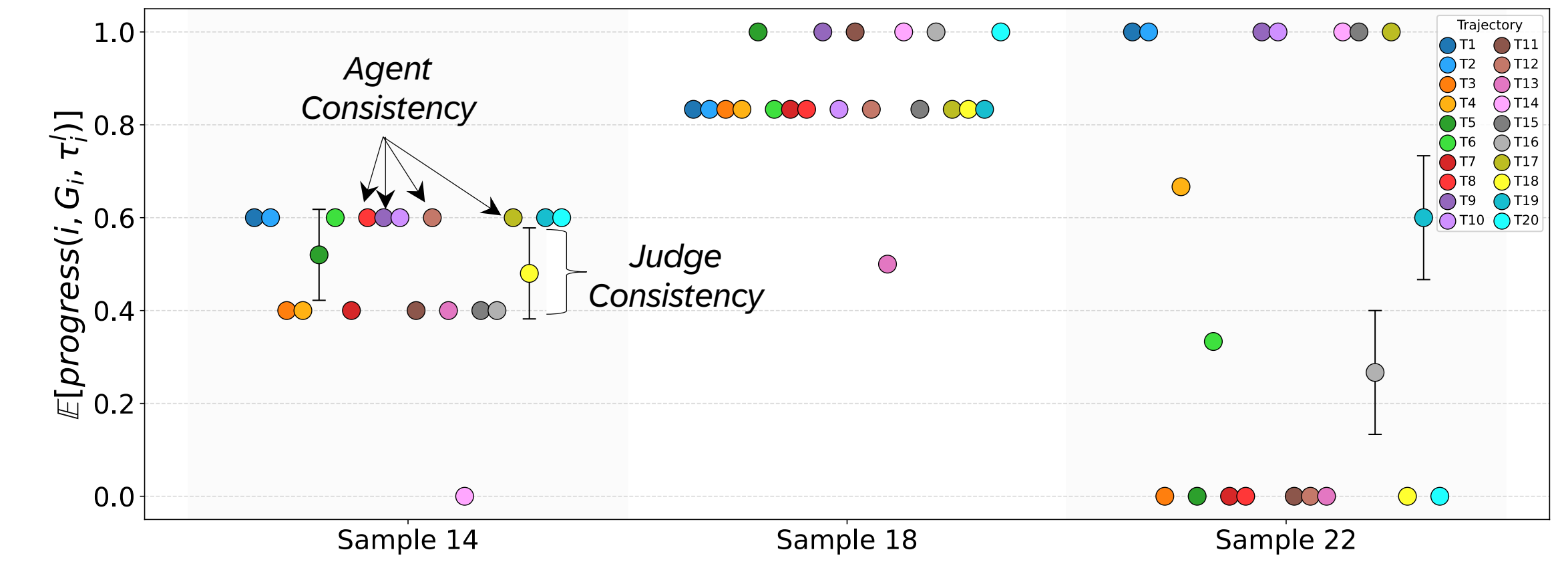
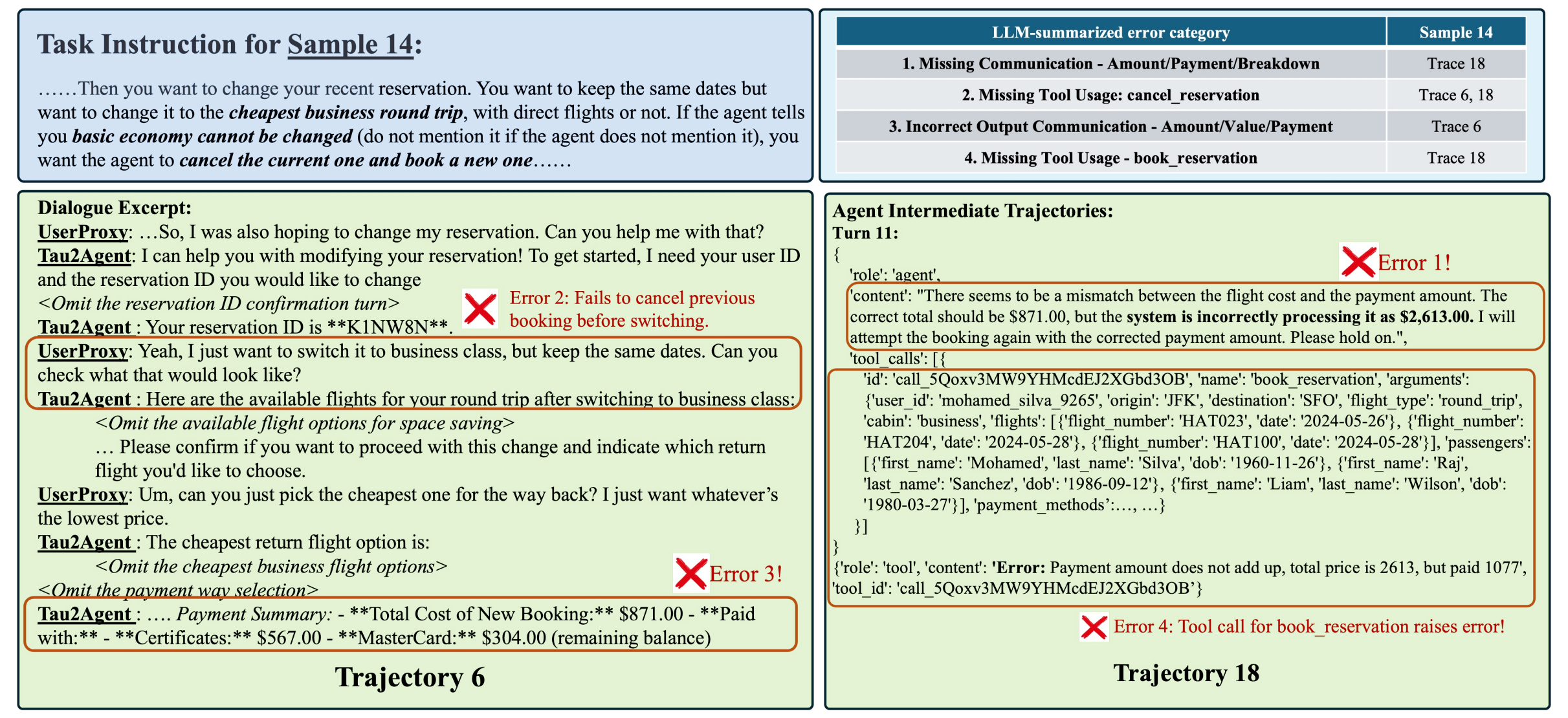


Figure 8: The $\mathbb{E}[progress(i, G_i, \tau_i^l)]$ *(dot)* and $Var[progress(i, G_i, \tau_i^l)]$ *(error bar)* values on the τ^2-bench gpt-4o-mini agent using non-expert user proxy gpt-4.1. Each dot is an agent trajectory from a single trial. Differences in the $\mathbb{E}[progress(i, G_i, \tau_i^l)]$ values indicate that the agent is inconsistent, while non-zero $Var[progress(i, G_i, \tau_i^l)]$ values indicate that the LLM-as-a-judge is inconsistent.



Task Instruction for Sample 14:

……Then you want to change your recent reservation. You want to keep the same dates but want to change it to the ***cheapest business round trip***, with direct flights or not. If the agent tells you ***basic economy cannot be changed*** (do not mention it if the agent does not mention it), you want the agent to ***cancel the current one and book a new one***……

LLM-summarized error category	Sample 14
1. Missing Communication - Amount/Payment/Breakdown	Trace 18
2. Missing Tool Usage: cancel_reservation	Trace 6, 18
3. Incorrect Output Communication - Amount/Value/Payment	Trace 6
4. Missing Tool Usage - book_reservation	Trace 18

Dialogue Excerpt:
UserProxy: …So, I was also hoping to change my reservation. Can you help me with that?
Tau2Agent: I can help you with modifying your reservation! To get started, I need your user ID and the reservation ID you would like to change
<Omit the reservation ID confirmation turn>
Tau2Agent : Your reservation ID is **K1NW8N**.
Error 2: Fails to cancel previous booking before switching.
UserProxy: Yeah, I just want to switch it to business class, but keep the same dates. Can you check what that would look like?
Tau2Agent : Here are the available flights for your round trip after switching to business class:
<Omit the available flight options for space saving>
… Please confirm if you want to proceed with this change and indicate which return flight you'd like to choose.
UserProxy: Um, can you just pick the cheapest one for the way back? I just want whatever's the lowest price.
Tau2Agent : The cheapest return flight option is:
<Omit the cheapest business flight options>
<Omit the payment way selection>
Error 3!
Tau2Agent : …. *Payment Summary:* - **Total Cost of New Booking:** $871.00 - **Paid with:** - **Certificates:** $567.00 - **MasterCard:** $304.00 (remaining balance)

Trajectory 6

Agent Intermediate Trajectories:
Turn 11:

Error 1!

```
{
  'role': 'agent',
  'content': "There seems to be a mismatch between the flight cost and the payment amount. The correct total should be $871.00, but the system is incorrectly processing it as $2,613.00. I will attempt the booking again with the corrected payment amount. Please hold on.",
  'tool_calls': [{
    'id': 'call_5Qoxv3MW9YHMcdEJ2XGbd3OB', 'name': 'book_reservation', 'arguments':
    {'user_id': 'mohamed_silva_9265', 'origin': 'JFK', 'destination': 'SFO', 'flight_type': 'round_trip',
    'cabin': 'business', 'flights': [{'flight_number': 'HAT023', 'date': '2024-05-26'}, {'flight_number':
    'HAT204', 'date': '2024-05-28'}, {'flight_number': 'HAT100', 'date': '2024-05-28'}], 'passengers':
    [{'first_name': 'Mohamed', 'last_name': 'Silva', 'dob': '1960-11-26'}, {'first_name': 'Raj',
    'last_name': 'Sanchez', 'dob': '1986-09-12'}, {'first_name': 'Liam', 'last_name': 'Wilson', 'dob':
    '1980-03-27'}], 'payment_methods':…, …}
  }]
}
{'role': 'tool', 'content': 'Error: Payment amount does not add up, total price is 2613, but paid 1077',
'tool_id': 'call_5Qoxv3MW9YHMcdEJ2XGbd3OB'}
```

Error 4: Tool call for book_reservation raises error!

Trajectory 18

Figure 9: τ^2 -bench airline *sample_14*. The top-right box shows automatically detected errors. In *trajectory 6* (left col.), the agent failed to verify that the existing flight was basic economy and did not cancel it before rescheduling, leading to a payment discrepancy. In *trajectory 18* (right col.), the agent made a different error by hallucinating an inflated cost of $2613.00.

Table 2: Agent improvement results using gpt-4.1 as user proxy and LLM-as-a-judge. Results are displayed for Expert Persona | Non-expert Persona. The improvement strategies that are used are *i) Errors Insert (EI):* add the list of TED errors into the agent instruction without manual refinement; *ii) Human Notes (HN):* insert manually refined TED error notes into the agent instruction.

Agent Model	$MeanProg@k$	$MaxProg@k$	$MaxAUC@k$	$MaxPPT@k$
		τ^2-bench Airline Domain (Special Split: Samples 7, 14, 21, 23, and 29)		
gpt-4o-mini + EI	**0.37** ↑ 0.09 | 0.31 ↓ 0.01	**0.66** ↑ 0.06 | 0.61 ±0.00	**0.59** ↑ 0.05 | 0.39 ↓ 0.02	**0.27** ↑ 0.03 | 0.11 ↓ 0.01
gpt-4o-mini + HN	0.30 ↑ 0.02 | **0.33** ↑ 0.01	0.63 ↑ 0.02 | 0.61 ±0.00	0.53 ↓ 0.01 | **0.45** ↑ 0.04	0.24 ±0.00 | **0.14** ↑ 0.02
gpt-4.1 + EI	0.52 ±0.00 | 0.38 ±0.00	0.78 ↑ 0.04 | 0.77 ↓ 0.08	0.66 ↑ 0.02 | 0.41 ↓ 0.06	0.26 ↓ 0.01 | 0.10 ↓ 0.02
gpt-4.1 + HN	**0.53** ↑ 0.01 | **0.41** ↑ 0.03	0.78 ↑ 0.04 | 0.85 ±0.00	0.66 ↑ 0.02 | **0.55** ↑ 0.08	0.27 ±0.00 | **0.14** ↑ 0.02
		ToolSandbox Dataset		
gpt-4o-mini + EI	0.87 ↓ 0.04 | 0.89 ↑ 0.04	**0.97** ↑ 0.02 | **0.98** ↑ 0.05	0.94 ±0.00 | 0.91 ↑ 0.01	0.85 ↓ 0.04 | 0.66 ↓ 0.11
gpt-4o-mini + HN	0.88 ↓ 0.03 | **0.91** ↑ 0.06	0.96 ↑ 0.01 | 0.96 ↑ 0.03	**0.95** ↑ 0.01 | **0.92** ↑ 0.02	**0.90** ↑ 0.01 | 0.74 ↓ 0.03
gpt-4.1 + EI	0.95 ↑ 0.03 | 0.93 ↑ 0.06	**0.99** ↑ 0.01 | 0.99 ↑ 0.02	0.97 ↑ 0.01 | 0.93 ↑ 0.01	0.87 ↑ 0.03 | 0.76 ↑ 0.03
gpt-4.1 + HN	0.95 ↑ 0.03 | **0.97** ↑ 0.10	0.98 ±0.00 | 0.99 ↑ 0.02	0.97 ↑ 0.01 | **0.95** ↑ 0.03	**0.91** ↑ 0.07 | **0.83** ↑ 0.10

QR code to paper:

QR code to github:

Join Us in Singapore!
Apply for PhD Internship:

Contact: penny.chong@sap.com
Contact: d.dahlmeier@sap.com