MTech Thesis Stage-1 Presentation

# LLMs for Travel Planning

**NLP**
**CS5803**
**(2024)**

Presented By :
**Shrenik Ganguli (CS23MTECH14014)**

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

# Contents

# Problem Statement

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- My Thesis Project delves into using LLMs for answering questions from the tourism domain.
- More specifically, its task would be to *create travel plans* based on a query given by the user.
- It would plan an entire itinerary from transportation, to accommodations, restaurants, tourist attractions etc.
- The language agent(s) would also have to adhere to all the *constraints* mentioned by the user in the query (such as budget, number of days).
- However, LLMs trained on Autoregressive decoding, do not do well for such planning tasks yet
- Therefore my focus is to understand in depth the problems faced by LLMs in planning tasks and then try to build a system where one or more language agents can handle the task of planning efficiently.

# Existing work in this domain

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- **TravelPlanner: A Benchmark for Real-World Planning with Language Agents [ICML 2024]**
  - Introduces a new *benchmark* called **TravelPlanner** to evaluate the performance of language agents in travel planning.
  - Introduces *constraints* that must be satisfied by a plan generated by it.
  - Introduces *evaluations metrics* to compare the generated plans.
  - Underscores the problems faced even by advanced LLMs like GPT-4 in making logical plans.

- **A Human-Like Reasoning Framework for Multi-Phases Planning Task with Large Language Models**
  - Improves on the TravelPlanner benchmark by developing a **human-like planning framework** for LLM agents.
  - Attempts to simulate various steps that humans take when planning such as generating an *outline, gathering necessary information and then proceeding to the planning task.*
  - Shows **10×** the performance gains in comparison to the baseline framework when deployed on GPT-4-Turbo.

- **Ask-before-Plan: Proactive Language Agents for Real-World Planning**
  - The above two papers assume that all the information needed for the planning task would be provided by the user which may *not* always be the case.
  - For this, a new task, **Proactive Agent Planning,** is introduced. This requires language agents to *predict clarification questions* based on user-agent conversation.
  - A novel multi-agent framework **Clarification-Execution-Planning (CEP),** is also proposed.

# Travel Planner

- TravelPlanner provides a rich sandbox environment with around four million data entries crawled from the Internet that can be accessed via six tools:

| Tool | Data Entries (#) |
|---|---|
| CitySearch | 312 |
| FlightSearch | 3,827,361 |
| DistanceMatrix | 17,603 |
| RestaurantSearch | 9,552 |
| AttractionSearch | 5,303 |
| AccommodationSearch | 5,064 |

- The training set includes 45 pairs of queries along with their human-annotated plans, the validation set includes 180 queries, and the test set includes 1,000 queries
- An example plan in the training set of TravelPlanner is given below:

```
[{'org': 'Las Vegas', 'dest': 'Denver', 'days': 3, 'visiting_city_number': 1, 'date': ['2022-03-12', '2022-03-13',
'2022-03-14'], 'people_number': 1, 'local_constraint': {'house rule': None, 'cuisine': None, 'room type': None,
'transportation': None}, 'budget': 1600, 'query': 'Please provide me with a travel plan that departs from Las Vegas to
Denver for 3 days, from March 12th to March 14th, 2022. The budget for this solo trip is $1,600.', 'level': 'easy'}, [
{'days': 1, 'current_city': 'from Las Vegas to Denver', 'transportation': 'Flight Number: F3629168, from Las Vegas to
Denver, Departure Time: 09:35, Arrival Time: 12:26', 'breakfast': '-', 'attraction': 'Denver Zoo, Denver;', 'lunch': 'The
Fatty Bao - Asian Gastro Bar, Denver', 'dinner': 'Nukkadwala, Denver', 'accommodation': 'Peaceful, beautiful home away ,
Denver'},
{'days': 2, 'current_city': 'Denver', 'transportation': '-', 'breakfast': 'Sweet Sensations, Denver', 'attraction': 'Beyond
Light Show, Denver;History Colorado Center, Denver;Big Blue Bear, Denver;', 'lunch': 'Yakooz, Denver', 'dinner': 'Kerala
Kitchen & Restaurant, Denver', 'accommodation': 'Peaceful, beautiful home away , Denver'},
{'days': 3, 'current_city': 'from Denver to Las Vegas', 'transportation': 'Flight Number: F3875298, from Denver to Las
Vegas, Departure Time: 20:59, Arrival Time: 21:47', 'breakfast': "Nathu's Sweets, Denver", 'attraction': 'Wings Over the
Rockies Air & Space Museum, Denver;', 'lunch': 'Tasty Fare, Denver', 'dinner': 'TBH - The Big House Cafe, Denver',
'accommodation': '-'}, {}, {}, {}, {}]]
```

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, Yu Su. "TravelPlanner: A Benchmark for Real-World Planning with Language Agents." ICML 2024

# Constraints



1. **Environment Constraints:** Constraints are manifested through the feedback received from the environment, assessing whether the language agent can adjust its plan appropriately.

| *Environment Constraint* | |
|---|---|
| Unavailable Transportation | There is no available flight or driving information between the two cities. |
| Unavailable Attractions | There is no available attraction information in the queried city. |

2. **Commonsense Constraints:** Constraints that are not explicitly mentioned to the language agent and must understand them by itself to follow common sense reasoning.

| *Commonsense Constraint* | |
|---|---|
| Within Sandbox | All information in the plan must be within the closed sandbox; otherwise, it will be considered a hallucination. |
| Complete Information | No key information should be left out of the plan, such as the lack of accommodation during travel. |
| Within Current City | All scheduled activities for the day must be located within that day's city(s). |
| Reasonable City Route | Changes in cities during the trip must be reasonable. |
| Diverse Restaurants | Restaurant choices should not be repeated throughout the trip. |
| Diverse Attractions | Attraction choices should not be repeated throughout the trip. |
| Non-conf. Transportation | Transportation choices within the trip must be reasonable. For example, having both "self-driving" and "flight" would be considered a conflict. |
| Minimum Nights Stay | The number of consecutive days spent in a specific accommodation during the trip must meet the corresponding required minimum number of nights' stay. |

3. **Hard Constraints:** Constraints specified by the user itself, such as budget constraints.

| *Hard Constraint* | |
|---|---|
| Budget | The total budget of the trip. |
| Room Rule | Room rules include "No parties", "No smoking", "No children under 10", "No pets", and "No visitors". |
| Room Type | Room types include "Entire Room", "Private Room", "Shared Room", and "No Shared Room". |
| Cuisine | Cuisines include "Chinese", "American", "Italian", "Mexican", "Indian", "Mediterranean", and "French". |
| Transportation | Transportation options include "No flight" and "No self-driving". |

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, Yu Su. "TravelPlanner: A Benchmark for Real-World Planning with Language Agents." ICML 2024

# Evaluation

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

The evaluation criteria includes the following metrics:

1) **Delivery Rate:** Assesses whether agents can successfully deliver a final plan within a limited number of steps (30 steps in this experimental setting).
2) **Commonsense Constraint Pass Rate:** Evaluates whether a language agent can incorporate common sense into their plan without explicit instructions.
3) **Hard Constraint Pass Rate:** Measures whether a plan satisfies all explicitly given hard constraints in the query.
4) **Final Pass Rate:** Represents the proportion of feasible plans that meet all aforementioned constraints among all tested plans.

For the Commonsense Constraint Pass Rate and Hard Constraint Pass Rate, two evaluation strategies, *micro* and *macro* are utilized.

- The **micro** strategy calculates the ratio of passed constraints to the total number of constraints

$$\text{Micro Pass Rate} = \frac{\sum_{p\in P} \sum_{c\in C_p} \mathbb{1}_{\text{passed}(c,p)}}{\sum_{p\in P} |C_p|},$$

- The **macro** strategy calculates the ratio of plans that pass all commonsense or hard constraints among all tested plans.

$$\text{Macro Pass Rate} = \frac{\sum_{p\in P} \mathbb{1}_{\text{passed}(C_p,p)}}{|P|}.$$

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, Yu Su. "TravelPlanner: A Benchmark for Real-World Planning with Language Agents." ICML 2024

# Experiments

- As can be seen from the table below, **GPT-4-Turbo** with **ReAct** achieves only **0.6%** in the final pass rate, and none of the other LLMs can pass any task

| | Validation (#180) | | | | | | Test (#1,000) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Delivery Rate | Commonsense Pass Rate | | Hard Constraint Pass Rate | | Final Pass Rate | Delivery Rate | Commonsense Pass Rate | | Hard Constraint Pass Rate | | Final Pass Rate |
| | | Micro | Macro | Micro | Macro | | | Micro | Macro | Micro | Macro | |
| Greedy Search | 100 | 74.4 | 0 | 60.8 | 37.8 | 0 | 100 | 72.0 | 0 | 52.4 | 31.8 | 0 |
| ***Two-stage*** | | | | | | | | | | | | |
| Mistral-7B-32K (Jiang et al., 2023) | 8.9 | 5.9 | 0 | 0 | 0 | 0 | 7.0 | 4.8 | 0 | 0 | 0 | 0 |
| Mixtral-8×7B-MoE (Jiang et al., 2024) | 49.4 | 30.0 | 0 | 1.2 | 0.6 | 0 | 51.2 | 32.2 | 0.2 | 0.7 | 0.4 | 0 |
| Gemini Pro (G Team et al., 2023) | 28.9 | 18.9 | 0 | 0.5 | 0.6 | 0 | 39.1 | 24.9 | 0 | 0.6 | 0.1 | 0 |
| GPT-3.5-Turbo (OpenAI, 2022) | 86.7 | 54.0 | 0 | 0 | 0 | 0 | 91.8 | 57.9 | 0 | 0.5 | 0.6 | 0 |
| GPT-4-Turbo (OpenAI, 2023) | **89.4** | **61.1** | **2.8** | **15.2** | **10.6** | **0.6** | **93.1** | **63.3** | **2.0** | **10.5** | **5.5** | **0.6** |
| ***Sole-planning*** | | | | | | | | | | | | |
| $\text{Direct}_{\text{GPT-3.5-Turbo}}$ | 100 | 60.2 | 4.4 | 11.0 | 2.8 | 0 | 100 | 59.5 | 2.7 | 9.5 | 4.4 | 0.6 |
| $\text{CoT}_{\text{GPT-3.5-Turbo}}$ | 100 | 66.3 | 3.3 | 11.9 | 5.0 | 0 | 100 | 64.4 | 2.3 | 9.8 | 3.8 | 0.4 |
| $\text{ReAct}_{\text{GPT-3.5-Turbo}}$ | 82.2 | 47.6 | 3.9 | 11.4 | 6.7 | 0.6 | 81.6 | 45.9 | 2.5 | 10.7 | 3.1 | 0.7 |
| $\text{Reflexion}_{\text{GPT-3.5-Turbo}}$ | 93.9 | 53.8 | 2.8 | 11.0 | 2.8 | 0 | 92.1 | 52.1 | 2.2 | 9.9 | 3.8 | 0.6 |
| $\text{Direct}_{\text{Mixtral-8x7B-MoE}}$ | 100 | 68.1 | 5.0 | 3.3 | 1.1 | 0 | 99.3 | 67.0 | 3.7 | 3.9 | 1.6 | 0.7 |
| $\text{Direct}_{\text{Gemini Pro}}$ | 93.9 | 65.0 | 8.3 | 9.3 | 4.4 | 0.6 | 93.7 | 64.7 | 7.9 | 10.6 | 4.7 | 2.1 |
| $\text{Direct}_{\text{GPT-4-Turbo}}$ | **100** | **80.4** | **17.2** | **47.1** | **22.2** | **4.4** | **100** | **80.6** | **15.2** | **44.3** | **23.1** | **4.4** |

- We can also see that agents *struggle in obtaining a high macro pass rate* although they achieve higher micro scores, indicating that although agents manage to satisfy some constraints, they often overlook some other constraints.

- Agents show a substantial gap between the two modes, indicating that they *struggle with both information collection and planning*. Alike humans, their performance declines when multitasking.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, Yu Su. "TravelPlanner: A Benchmark for Real-World Planning with Language Agents." ICML 2024

# Problems identified

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- We see in the first Tool-use example, agent fails to complete a plan due to repeated use of incorrect dates. This shows that current agents *cannot self-correct their initial mistakes*.
- In the Direct example, agent mistakenly uses the same flight number for both departure and return flights, *resulting in hallucinations*, and suggest that agents might be *lost when faced with mass information*.
- As illustrated in the Reflexion example, despite recognizing the necessity to minimize costs, agents tend to randomly select items, some of which are more expensive, thus *not aligning their actions with their reasoning.*

**Tool-use**

**Query:** Please plan a travel itinerary for me. I'm departing from Cincinnati and heading to Norfolk for three days. The dates of travel are from March 10th to **March 12th, 2022**. I have a budget of $1,400 for this trip.

**Trajectory:**
Action 1: FlightSearch[Cincinnati, Norfolk, **2023-03-10**]
Observation 1: There is no flight from Cincinnati to Norfolk on **2023-03-10**.
Action 2: FlightSearch[Cincinnati, Norfolk, **2023-03-09**]
Action 3: FlightSearch[Cincinnati, Norfolk, **2023-03-11**]
Action 4: FlightSearch[Cincinnati, Richmond, **2023-03-10**]
...
Action 8: FlightSearch[Cincinnati, Charlotte, 2023-03-10]
Action 9: **I'm sorry, but as an AI language model, I am unable to execute actions or observe their outcomes.**

**Direct**

**Query:** Please create a travel itinerary for a solo traveler departing from Jacksonville and heading to Los Angeles for a period of 3 days, from March 25th to March 27th, 2022. The budget for this trip is now set at $2,400.

**Plan:**
Day 1:
transportation: **Flight Number: F3765485, from Jacksonville to Los Angeles**
...
accommodation: Midtown East 2 Bed/2 Bath Elevator Building Apt (7-day minimum night)

Day 2:
**...**
accommodation: Midtown East 2 Bed/2 Bath Elevator Building Apt (7-day minimum night)

Day 3:
transportation: **Flight Number: F3765485, from Los Angeles to Jacksonville**
**...**

**Reflexion**

**Query:** Can you help construct a travel plan that begins in Philadelphia and includes visits to 3 cities in Virginia? The trip duration is for 7 days, from March 15th to March 21st, 2022, with a total budget of $1,800.

**Trajectory:**
...
Observation 3: **The cost of your current plan is 316.0 dollars.**

Thought 4: **I need to select a more budget-friendly accommodation** to keep the overall trip within the $1,800 budget.
Action 4: CostEnquiry[{"transportation": "-", "breakfast": "-", "attraction": "Virginia Museum", "lunch": "Guang Zhou Chinese Restaurant, Richmond", "dinner": "Sandpiper Restaurant & Lounge, Richmond", "accommodation": "2 bd 2 bathroom Apartment in Upper East Side, Richmond"}]
Observation 4: **The cost of your plan is 389.0 dollars.**

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, Yu Su. "TravelPlanner: A Benchmark for Real-World Planning with Language Agents." ICML 2024

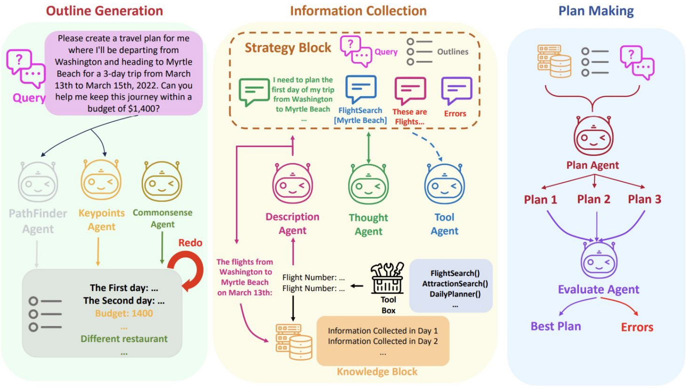
# Human-like reasoning framework

- This method aims to improve on the TravelPlanner benchmark by developing a human-like planning framework for LLM agents for solving *Multi-Phases* problems.
- They implement several strategies to enable LLM agents to generate a coherent **outline** for each travel query, mirroring human planning patterns.
- Additionally, they integrate **Strategy Block** and **Knowledge Block** into our framework: Strategy Block facilitates information collection, while Knowledge Block stores and provides essential information for detailed planning.
- Then in the final phase, **Plan Search** is implemented, where multiple potential plans for the same travel query are generated and then the most suitable plan is selected by the **Evaluation Agent**.
- Such strategies use *multi-agent frameworks*, demonstrating superior performance compared to single-agent systems in solving diverse tasks.

Chengxing Xie, Difan Zou. "A Human-Like Reasoning Framework for Multi-Phases Planning Task with Large Language Models"

# Phases of the framework

- **Outline Generation Phase:** The route is generated by **PathFinder Agent**, which is then evaluated from a transportation and common sense perspective, to be then provided to the subsequent phases.
- **Information Collection Phase:** The outline generated is first stored in the **Strategy Block**, which then informs the **Thought Agent** which day it is and short descriptions of the collected data. The Thought Agent generates the next steps based on this information, whose output is then utilized by the **Tool Agent** to generate a suitable function expression. The result of this function, primarily comprising of travel information, is then recorded in the **Knowledge Block** with a description from the **Description Agent**. Finally, the selected information is forwarded to the Plan Agent**.**
- **Plan Making Phase:** Armed with sufficient information, the **Plan Agent** along with the **DailyPlanner** tool, proceed to formulate a plan, typically for a specific day, using Plan Search.






Chengxing Xie, Difan Zou. “A Human-Like Reasoning Framework for Multi-Phases Planning Task with Large Language Models”

# Results

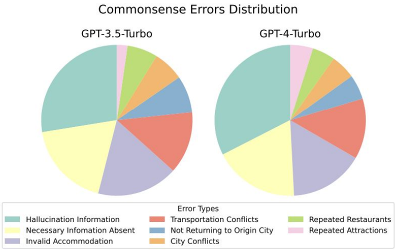
- Results demonstrate that the framework increases the Final Pass Rate of GPT-4-Turbo by **10×** compared to previous algorithms.

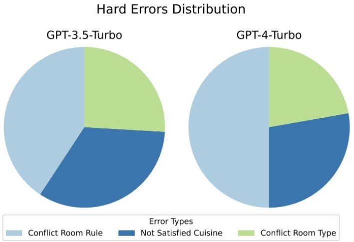
| Model | Delivery Rate | Commonsense Pass Rate | | Hard Constraint Pass Rate | | Final Pass Rate |
|---|---|---|---|---|---|---|
| | | Micro | Macro | Micro | Macro | |
| ***TravelPlanner Result*** | | | | | | |
| Mistral-7B-32K [21] | 8.9 | 5.9 | 0 | 0 | 0 | 0 |
| Mixtral-8×7B-MoE [22] | 49.4 | 30.0 | 0 | 1.2 | 0.6 | 0 |
| GPT-3.5-Turbo [34] | 86.7 | 54.0 | 0 | 0 | 0 | 0 |
| GPT-4-Turbo [1] | 89.4 | 61.1 | 2.8 | 15.2 | 10.6 | 0.6 |
| ***Our Results*** | | | | | | |
| Mistral-7B-32K [21] | 39.4 | 24.0 | 1.1 | 0.5 | 0.6 | 0.6 |
| Mixtral-8×7B-MoE [22] | 67.8 | 40.2 | 0 | 0 | 0 | 0 |
| GPT-3.5-Turbo [34] | **100.0** | **75.1** | 15.6 | 15.5 | 4.4 | 2.2 |
| GPT-4-Turbo [1] | 91.7 | 74.6 | **24.4** | **35.7** | **16.7** | **6.7** |

- To see the effect of each component of the framework

| | Delivery Rate | Commonsense Pass Rate | | Hard Constraint Pass Rate | | Final Pass Rate |
|---|---|---|---|---|---|---|
| | | Micro | Macro | Micro | Macro | |
| Original Result [54] | 86.7 | 54.0 | 0 | 0 | 0 | 0 |
| No Outlines | 93.3 | 59.2 | 1.1 | 3.8 | 1.1 | 0 |
| No Strategy | 100 | 74.9 | 13.9 | 12.1 | 2.2 | 0.6 |
| No Knowledge | 99.4 | 66.0 | 8.3 | 10.2 | 1.1 | 0 |
| No Plan Search | 99.4 | 67.4 | 0.6 | 1.9 | 0 | 0 |
| Our Framework | **100.0** | **75.1** | **15.6** | **15.5** | **4.4** | **2.2** |

- The error distribution of GPT-3.5-Turbo and GPT-4-Turbo.




Chengxing Xie, Difan Zou. “A Human-Like Reasoning Framework for Multi-Phases Planning Task with Large Language Models”

# Ask-before-Plan

- To study the practical problem of a user not specifying all the necessary details in the query, this paper introduces a new task, which is **Proactive Agent Planning.**
- Proactive Agent Planning requires language agents to *predict clarification needs* based on user-agent conversation and agent-environment interaction, invoke external tools to collect valid information, and finally generate a plan to fulfill the user's demands.
- A new benchmark dataset , **Ask-before-Plan,** is also established.
- They propose a novel multi-agent framework, **Clarification-Execution-Planning (CEP)**, which consists of three agents specialized in clarification, execution, and planning:
    - The **Clarification agent** is responsible for understanding the uncertainty of user instructions and asking users clarifying questions to uncover their underlying intentions.
    - The **Execution agent** leverages various tools to interact with the environment, gathering necessary information for the Clarification agent.
    - The **Planning agent** produces the final plan by aggregating the clarification process for accomplishing the initial user instruction.

Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, Tat-Seng Chua. Ask-before-Plan: Proactive Language Agents for Real-World Planning.

# Ask-before-Plan Dataset

| org | dest | days | departure_ date | people_ number | budget | query | level | details |
|---|---|---|---|---|---|---|---|---|
| Las Vegas | null | null | 2022-03-12 | 1 | null | I am planning a solo trip departing from Las Vegas on March 12th,2022. | easy | [{'operation': 'ADD', 'attribute': 'budget', 'value': 1600, 'priority': 0, 'question': 'Do you have a specific budget in mind for your solo trip departing from Las Vegas on March 12th, 2022?', 'answer': 'My budget for the trip is $1600.'},<br><br>{'operation': 'ADD', 'attribute': 'days', 'value': 3, 'priority': 0, 'question': 'How many days are you planning to spend on your solo trip departing from Las Vegas on March 12th, 2022?', 'answer': "I'm planning to spend 3 days on my solo trip departing from Las Vegas on March 12th, 2022."},<br><br>{'operation': 'ADD', 'attribute': 'dest', 'value': ['Denver on day 1'], 'priority': 0, 'question': 'Which cities are you considering visiting during your trip, and on which specific days do you plan to arrive in each city?', 'answer': "Oh, I can't wait to explore Denver on the first day of my trip! The city has so much to offer, from beautiful parks to vibrant cultural attractions. I'm already looking forward to immersing myself in everything it has to offer."}] |

Taken from Ask-before-Plan Hugging Face

# Methodology

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

The method outlines the approach taken to address the problem of Proactive Agent Planning. It includes several key components:

- **Trajectory Tuning:** It is introduced to fine-tune the agents. The trajectory tuning scheme uses the user-agent *conversation history (Ct-1)* and agent-environment *interaction history (Et-1)* to fine-tune the agents.

- **Clarification Agent:** The clarification agent operates in two steps:
    - **Clarification Need Prediction**: The agent generates a *boolean indicator ( b_t )* to predict the need for clarification based on the ongoing conversation (Ct-1) and the last turn interaction (Et-1).
    - **Clarification Question Generation**: If ( b_t ) is true, the model generates a *clarifying question ( a_t )* for a *specific detail ( d_t )*. The conversation sequence is tokenized into *input tokens (x_t)* using a prompt template function.

- **Training Objective:** This involves learning the model parameters over the total number of conversation turns and the number of tokens in each turn.

$$\mathcal{L} = \max_{\theta} \frac{1}{|T|} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \log P_{\theta}^{t}(x_i^t | \mathbf{x}_{<i}^t)$$



Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, Tat-Seng Chua. Ask-before-Plan: Proactive Language Agents for Real-World Planning.

# Results



| | Clarif. Acc | | Rule-based Score | | BLEU | GPT Score | |
|---|---|---|---|---|---|---|---|
| | Micro | Macro | Micro | Macro | | Micro | Macro |
| Environment-only | 70.4 | 17.7 | 21.5 | 8.1 | 1.0 | 40.1 | 19.4 |
| **Conversation-only** | | | | | | | |
| Proactive (GPT-3.5) | 62.3 | 6.1 | 9.7 | 3.4 | 3.7 | 0.9 | 0 |
| ProCoT (GPT-3.5) | 33.7 | 10.6 | 3.3 | 2.6 | 2.2 | 2.4 | 1.8 |
| **Environment + Conversation** | | | | | | | |
| Direct (GPT-3.5) | 47.0 | 16.9 | 20.8 | 17.4 | 8.2 | 8.6 | 6.2 |
| CEP (Mistral-7B) | 82.8 | 51.7 | 54.2 | 37.0 | 44.5 | 73.1 | 58.6 |
| CEP (LLaMA-3-8B) | **99.4** | **98.2** | **69.7** | **55.8** | **57.2** | **85.8** | **77.0** |

**Table for Evaluation of Clarification**: As can be seen, the results of the environment+conversation method surpass those of the conversation-only and environment-only methods, pointing out the importance of using environment observation as a supervised signal in clarification for language agents.

| | Delivery Rate | Commonsense Pass Rate | | Hard Constraint Pass Rate | | Final Pass Rate |
|---|---|---|---|---|---|---|
| | | Micro | Macro | Micro | Macro | |
| Greedy Search | 100 | 76.9 | 0 | 64.5 | 46.7 | 0 |
| Direct (Mistral-7B) | 86.6 | 44.8 | 0.4 | 4.0 | 0.9 | 0 |
| CoT (Mistral-7B) | 61.5 | 29.8 | 0 | 2.4 | 0.1 | 0 |
| Direct (GPT-3.5) | **98.6** | **63.7** | **0.7** | **19.0** | **5.1** | **0.1** |
| CoT (GPT-3.5) | 77.5 | 50.0 | 0.6 | 16.2 | 5.2 | 0 |
| ReAct (GPT-3.5) | 68.7 | 38.0 | 0 | 3.2 | 0.6 | 0 |
| Reflexion (GPT-3.5) | 61.5 | 33.9 | 0 | 3.1 | 0.4 | 0 |

**Table for Evaluation of Planning:**

- For Mistral, we only test the Direct and CoT strategies because the agent gets trapped in a fail-retry loop when using ReAct and Reflexion.
- For GPT3.5, we observe the same trend in TravelPlanner, the performance degrades from direct prompting to Reflexion, demonstrating the difficulty for most advanced agents to finish the planning task.

| **Clarification** | | | | | | |
|---|---|---|---|---|---|---|
| | Clarif. Acc | | Rule-based Score | | GPT Score | |
| | Micro | Macro | Micro | Macro | Micro | Macro |
| $\text{CEP}_{\text{independent}}$ (LLaMA-3-8B) | 99.4 | 98.2 | 69.7 | 55.8 | 85.8 | 77.0 |
| $\text{CEP}_{\text{integral}}$ (LLaMA-3-8B) | 97.3 | 92.9 | 68.4 | 54.7 | 85.1 | 76.0 |
| **Planning** | | | | | | |
| | Delivery Rate | Commonsense Pass Rate | | Hard Constraint Pass Rate | | Final Pass Rate |
| | | Micro | Macro | Micro | Macro | |
| $\text{CEP}_{\text{integral}}$ | 98.8 | 64.3 | 1.0 | 19.2 | 5.0 | 0.1 |
| $\text{CEP}_{\text{integral}}$ w/o Clarification | 93.3 | 53.3 | 0.3 | 8.4 | 3.1 | 0 |

**Table for Evaluation of the integral framework:** The evaluation uses the interactions generated by the state-of-the-art (SOTA) model, CEP (LLaMA-3-8B), for subsequent clarification and final planning. Results indicate that the integral framework, including clarification and execution interactions, performs robustly, although there is a performance drop when it comes to satisfying all the constraints indicated by dismal final pass rates.

Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, Tat-Seng Chua. Ask-before-Plan: Proactive Language Agents for Real-World Planning.

# Case Studies

- The two case studies given below explore the impact of clarification on subsequent planning subtasks.
- In the first case, it is evident that without asking for the missing destination detail, the agent incorrectly confines the itinerary to the origin city, Nashville, on the first day, leading to a confusing itinerary.
- In the second case if the agent overlooks that the previously specified accommodation preferences in Houston are unavailable and fails to inquire about the user's opinion, an incomplete travel plan will be generated.

**Instruction**: Could you suggest a 5-day travel itinerary for a solo traveler starting from Nashville, departing on March 13th, 2022, with a budget of $2123?
**Missing Detail**: Destination by arrival days.
**Question**: Which cities are you interested in visiting during your 5-day trip, and on which specific days do you plan to arrive in each city?
**Response**: I can't wait to explore the vibrant streets of San Diego on day 1 and then soak up the glitz and glamour of Los Angeles on day 3!

| | $\mathrm{CEP}_{\mathrm{integral}}$ | $\mathrm{CEP}_{\mathrm{integral}}$ w/o Clarification |
|---|---|---|
| **Day** | 1 | 1 |
| **Current City** | from Nashville to San Diego | Nashville |
| **Transportation** | Flight Number: F3973475, from Nashville to San Diego, Departure Time: 09:14, Arrival Time: 11:22, Cost: $485 | N/A |
| **Breakfast** | Open Yard, San Diego | Biscuit Love Gulch, Nashville |
| **Attraction** | Cabrillo National Monument, San Diego; La Jolla Shores Park, San Diego | Country Music Hall of Fame and Museum, Nashville; Ryman Auditorium, Nashville |
| **Lunch** | Bikaner Sweets, San Diego | Hattie B's Hot Chicken, Nashville |
| **Dinner** | Chawlas 2, San Diego | Martin's Bar-B-Que Joint, Nashville |
| **Accommodation** | Spacious Room in Large 2 Bedroom Prewar Apartment, San Diego | Cozy Loft in East Nashville, Nashville |

**Instruction**: I am planning a solo trip starting from Bloomington. The itinerary includes visiting Dallas on day 1, Amarillo on day 3, and Houston on day 5. I will be departing on March 5th, 2022. I am looking for shared room accommodations that allow pets.
**Unfeasible Detail**: Accommodation.
**Question**: It seems like the accommodation preferences you've specified might not be feasible for this trip. Is there a different type of accommodation you would consider for your solo trip?
**Response**: Any accommodation preference is acceptable.

| | $\mathrm{CEP}_{\mathrm{integral}}$ | $\mathrm{CEP}_{\mathrm{integral}}$ w/o Clarification |
|---|---|---|
| **Day** | 5 | 5 |
| **Current City** | from Amarillo to Houston | from Amarillo to Houston |
| **Transportation** | Flight Number: F3822285, from Amarillo to Houston, Departure Time: 06:31, Arrival Time: 08:05, Cost: $103 | Taxi: from Amarillo to Houston, Duration: 8 hours 49 mins, Distance: 965 km, Cost: 965 |
| **Breakfast** | N/A | Jalapenos, Houston |
| **Attraction** | Downtown Aquarium, Houston; Space Center Houston, Houston | Downtown Aquarium, Houston; Space Center Houston, Houston |
| **Lunch** | Jalapenos, Houston | The BrewMaster - The Mix Fine Dine, Houston |
| **Dinner** | Royal Mart, Houston | "Pebble Street, Houston |
| **Accommodation** | Superhost 3 bedroom DISCOUNT, Houston | N/A |

Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, Tat-Seng Chua. Ask-before-Plan: Proactive Language Agents for Real-World Planning.

# Future Work

- The literature survey I have done till now has given me a deeper understanding of the problems faced by LLMs in the planning task and my future work in this domain would be to improve on these aspects to optimize the planning.
- In particular, I will be working to improve their performance in tool usage by using existing technologies like **Gorilla LLM** to invoke semantically and syntactically correct API sequences.
- Another area I would be working is to try and replace some language agents (in multi-agent framework) by **traditional deep learning algorithms/models** since it would be easier to optimise their performance.
- I would also to improve on the fine tuning methods used by these papers, like using **Parameter Efficient Fine-Tuning** methods like LoRA and QLoRA to see if they can provide performance boosts.
- At the end, would like to acknowledge, there is a lot of hit and trial that must be done for overall optimization of any ML task and this would be no different. Hence, that would be the goal for the next stages my thesis project.

# References

[1] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, Yu Su. TravelPlanner: A Benchmark for Real-World Planning with Language Agents. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[2] Xie, Jian, et al. "TRAVELPLANNER: A Benchmark for Real-World Planning with Language Agents." In Proceedings of the International Conference on Learning Representations (ICLR), 2024.

[3] Chengxing Xie, Difan Zou. A Human-Like Reasoning Framework for Multi-Phases Planning Task with Large Language Models. arXiv preprint arXiv:2405.18208, 2024

[4] Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, Tat-Seng Chua. Ask-before-Plan: Proactive Language Agents for Real-World Planning. arXiv preprint arXiv:2406.12639, 2024

[5] Shishir G. Patil, Tianjun Zhang, Xin Wang, Joseph E. Gonzalez. Gorilla: Large Language Model Connected with Massive APIs. arXiv preprint arXiv:2305.15334, 2023

# THANK YOU!